State - Space models for longitudinal data

Veronica Boero Rodriguez

June, 2004

A thesis submitted for a degree of Master of Philosophy in Statistics of the Australian National University.

Certificate of Originality

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no materials previously published or written by another person nor material which to a substantial extent has been accepted for the award of any other degree or diploma of a university or other institute of higher education, except where due acknowledgement is made in the text of the thesis.

Veronica Boero Rodriguez

Acknowledgements

I would like to specially thank my supervisor Alan Welsh for all the help, encouragement, advice, support and patience. Specially his patience over all these years.

I am very grateful to my colleagues at the Australian Bureau of Agricultural and Resources Economics (ABARE) for their support and encouragement. In particular, Ray Lindsay, who helped me with the SAS programming and Greg Griffiths, for proofreading and improving this thesis.

Thanks to Ann Cowling from the Statistical Consultancy Unit at the Australian National University, for providing me the diabetes data.

And finally, but not least, I am very grateful to my family. To my husband Hector and sons Facundo and Nicholas for their support throughout these years, I could not have done it without it. To my parents, especially to my mother for her support and encouragement when I was feeling like giving up.

Thank you to you all,

Veronica

Abstract

State-space methods and the Kalman filter are powerful tools to handle models that contain error correlation structure. They use recursive calculations with small matrices, the size of which do not depend on the number of observations available on the subject.

In this thesis, the general form of state-space models and the Kalman filter are studied. Some examples are also presented in state-space form, starting with a simple AR(1) model and progressing to more complex models like the continuous AR (1), simple linear regression and the Laird-Ware model. These models are written in state-space form and the Kalman filter is applied to them to obtain their likelihood function.

Usually the aim of obtaining the likelihood function of a model is to estimate its parameters using maximum likelihood estimation. To do so, it is necessary to obtain the derivatives of the likelihood function. We show how to obtain these derivatives using the Kalman filter.

To better illustrate these methods, two applications are also presented. The growth of Sitka spruce trees data, Diggle et al (1994) and the diabetes data. A Laird-Ware model with no observational error and an AR(1) error correlation structure is fitted to the Sitka spruce data. A similar model is fitted to the diabetes data but with a CAR(1) error correlation structure. Both examples showed that the error correlation structure cannot be ignored. It is also shown in the diabetes example that by using state-space models and the Kalman filter, we avoid inverting matrices of dimensions up to 285×285.

We also show that even though there are statistical packages that have functions avail-

able to fit the Laird-Ware model when the errors have a discrete correlation structure, there are none for when they have a continuous correlation structure. State-space models and the Kalman filter allow an easy way to fit these much more flexible and complicated models.

Contents

1 Introduction 4

2 The Kalman Recursion 8

2.1 The state-space equations 8

2.2 The Kalman recursion 10

2.3 Using the Kalman recursion to calculate the likelihood function of an AR(1) process 12

2.4 Unequally spaced data 18

2.5 Using the Kalman recursion to calculate the likelihood function of a CAR(1) process 21

3 The linear regression model 24

3.1 Calculating the likelihood using the Cholesky decomposition 24

3.2 Calculating the likelihood using the Kalman Filter 26

3.3 The Laird - Ware model 36

3.4 The Laird-Ware model in state-space form 37

4 Using the Kalman filter to obtain the maximum likelihood estimators of the parameters **40**

4.1 The likelihood function 40

4.2 Newton-Raphson method 42

4.3 Fisher's method of scoring 43

4.4 Using the Kalman filter to obtain the derivatives and information matrix of the log-likelihood function 44

5 Applications **50**

5.1 Growth of Sitka spruce data 50

5.2 Diabetes data 56

6 Conclusion **68**

Chapter 1

Introduction

Longitudinal data analysis is the analysis of data collected on subjects which are followed over time. Because there are multiple observations on each subject, these observations tend to be correlated, and this correlation must be accounted for in order to produce a proper analysis.

One method used to estimate model parameters is maximum likelihood. In most cases, maximum likelihood estimation requires searching over the parameter space in an attempt to find the values of the unknown parameters that maximize the likelihood function. Since algorithms for non linear optimization are readily available, the difficult part often is to express the likelihood in a form that can be calculated. When the data from each subject are serially correlated, the state-space approach provides a convenient way to compute likelihoods using the Kalman filter.

Kalman (1960) developed an approach to filtering and forecasting based on the concept of state space transition. State - space methods use recursive calculations with small

matrices, the size of which do not depend on how many observations are available on the subject. This methodology was used by the aerospace industry for estimating the position of rockets. It is a real time estimation procedure that allows the estimates of the rocket's position to be continually updated as the data are collected. The term filter is used because the estimation of the position and velocity of the rocket, the state of the system, is carried out at the time of the most recent set of observations. Since the observations have random errors, estimates are obtained by using information about where the rocket was estimated to be at the previous time and where it was predicted to be at the present time. However, Kalman's methodology, as originally developed, does not estimate any parameters, variances and other parameters in the model are assumed to be known.

Later on, it was shown that when a problem can be set up in state - space form, the methodology exists for constructing the likelihood. Schweppe (1965) showed that if errors have a Gaussian distribution, the Kalman filter could be used to calculate likelihoods. The power of the Kalman approach to calculating likelihoods is that when a problem can be formulated in state - space form, it is possible to calculate the likelihood recursively without inverting large matrices.

Laird and Ware (1982) proposed a very general linear mixed model for longitudinal data. The calculation of likelihoods for estimating the parameters of this type of model requires inverting a covariance matrix of size n x n where n is the number of observations on a subject. There is no problem when n is small, but these computations are troublesome when n is large.

The Laird-Ware model can be set up in state - space form. The Kalman filter can then

be used to calculate the exact likelihood and nonlinear optimization used to obtain the maximum likelihood estimates. In this way, it is possible to analyze complex models, and to do so, it is not necessary to invert large covariance matrices.

Chapter 2 shows the general form of a state - space model. Linear models can be written in state - space form and the Kalman filter can be used to write their likelihood function. This chapter also includes examples on how autoregressive and continuous autoregressive models can be written in state - space form and their likelihood calculated.

In chapter 3 a linear regression model is written in state - space form. It is shown how the likelihood function of this model can be calculated using the Kalman filter. Also in this chapter the Laird - Ware model is introduced, and we show how its likelihood function can be calculated using the Kalman filter.

Usually the aim of obtaining the likelihood function of a model is to estimate its parameters using maximum likelihood estimation. To do so, it is necessary to obtain the derivatives of the likelihood function. Chapter 4 shows how the derivatives of the likelihood function can be calculated using the Kalman filter.

Chapter 5 shows two applications of state - space models for longitudinal data. Part of the spruce data in Diggle et al (1994) was analyzed in this section as a simple example of a regression model with correlated errors in state - space form. The second example is the diabetes data. For this example, a model with a more complicated error structure was considered.

Finally, the last chapter includes a discussion about the state-space models analyzed in this thesis and the advantages of using the Kalman filter when fitting the more complicated models.

Chapter 2

The Kalman Recursion

2.1 The state-space equations

Suppose that $t_1 < t_2 < \ldots < t_n$ are a set of time points and that there is a n-vector of observations $\{y(t_1), y(t_2), \ldots, y(t_n)\}$. A state-space model relates the observed n-vector $\mathbf{y}(t_i)$ to an unobserved m-vector state $\mathbf{s}(t_i)$ which is assumed to evolve in time. The relationship between $\mathbf{y}(t_i)$ and $\mathbf{s}(t_i)$ is given by the *measurement equation*

$$\mathbf{y}(t_i) = \mathbf{H}(t_i)\mathbf{s}(t_i) + \mathbf{\Sigma}(t_i)\mathbf{v}(t_i), \tag{2.1}$$

where $\mathbf{H}(t_i)$ is a $n \times m$ matrix and $\mathbf{v}(t_i)$ is the observational error vector which is assumed to have a Normal distribution with zero mean and covariance matrix $\mathbf{I}$, and $\mathbf{\Sigma}(t_i)$ is the $n \times n$ scaling covariance matrix.

The matrix $\mathbf{H}(t_i)$ shows which elements of the state vector are observed, or which linear combinations of the state vector are observed. The matrix $\mathbf{H}(t_i)$ can be differ-

ent at different time points. A more general formulation includes a n-vector $\mathbf{f_0}(t_i)$ of non-random inputs; however, in most problems $\mathbf{f_0}(t_i) = 0$.

In general, the elements of $\mathbf{s}(t_i)$ are not observable. However, they are assumed to be generated by a first order Markov process, which is described by the *transition equation*

$$\mathbf{s}(t_i) = \mathbf{\Phi}(t_i, t_{i-1})\mathbf{s}(t_{i-1}) + \mathbf{f_s}(t_i) + \mathbf{\Psi}(t_i)\mathbf{u}(t_i), \tag{2.2}$$

where $\mathbf{\Phi}(t_i, t_{i-1})$ is the $m \times m$ state transition matrix from time t_{i-1} to time t_i, $\mathbf{f_s}(t_i)$ is an m-vector of non-random inputs, $\mathbf{u}(t_i)$ is the random input to the state at time (t_i) which is assumed to be normally distributed with mean zero and covariance $\mathbf{I}$ and independent of the observational error, and $\mathbf{\Psi}(t_i)$ is the scaling covariance matrix.

The variance of the $\mathbf{\Sigma}(t_i)\mathbf{v}(t_i)$ is $\mathbf{\Sigma}(t_i)\mathbf{\Sigma}(t_i)^T$. The covariance matrix of the random input to the transition equation over a time interval $(t_i - t_{i-1})$ is $\mathbf{\Psi}(t_i)\mathbf{\Psi}(t_i)^T$. The disturbances $\mathbf{u}(t_i)$ and $\mathbf{v}(t_i)$ are uncorrelated with each other in all time periods, and uncorrelated with the initial state.

The specification of the state-space system is completed by specifying that the initial state vector, $\mathbf{s}(0|0)$, has mean a_0 and covariance matrix $\mathbf{P}(0|0)$. The initial conditions for the Kalman recursion are given by setting these to the values of the mean and covariance matrix of the unconditional distribution of the state vector respectively.

Usually, the observations $\mathbf{y}(t_i)$ and the matrix $\mathbf{H}(t_i)$ are known and the parameters in $\mathbf{\Phi}(t_i, t_{i-1})$, in the covariance matrix $\mathbf{\Sigma}(t_i)$ and in the covariance matrix $\mathbf{\Psi}(t_i)$ are to be estimated.

2.2 The Kalman recursion

The recursion starts with the estimate of the state at time t_{i-1}, $\mathbf{s}(t_{i-1}|t_{i-1})$ and its covariance matrix $\mathbf{P}(t_{i-1}|t_{i-1})$.

The Kalman recursion steps are:

1. Calculate a one step prediction of the state at time t_i

$$\mathbf{s}(t_i|t_{i-1}) = \mathbf{\Phi}(t_i, t_{i-1})\mathbf{s}(t_{i-1}|t_{i-1}) + \mathbf{f_s}(t_i) \tag{2.3}$$

and calculate the covariance matrix of the prediction,

$$\mathbf{P}(t_i|t_{i-1}) = \mathbf{\Phi}(t_i, t_{i-1})\mathbf{P}(t_{i-1}|t_{i-1})\mathbf{\Phi}(t_i, t_{i-1})^T + \mathbf{\Psi}(t_i)\mathbf{\Psi}(t_i)^T. \tag{2.4}$$

2. Calculate the prediction of the next observation vector,

$$\mathbf{y}(t_i|t_{i-1}) = \mathbf{H}(t_i)\mathbf{s}(t_i|t_{i-1}) \tag{2.5}$$

and calculate the innovation vector which is the difference between the observation vector and the predicted observation vector,

$$\mathbf{e}(t_i) = \mathbf{y}(t_i) - \mathbf{y}(t_i|t_{i-1}). \tag{2.6}$$

3. Calculate the covariance matrix of the innovation vector $\mathbf{e}(t_i)$,

$$\mathbf{V}(t_i) = \mathbf{H}(t_i)\mathbf{P}(t_i|t_{i-1})\mathbf{H}(t_i)^T + \mathbf{\Sigma}(t_i)\mathbf{\Sigma}(t_i)^T. \tag{2.7}$$

4. Accumulate the quantities needed to calculate the log-likelihood at the end of the recursion,

$$RSS \leftarrow RSS + \mathbf{e}(t_i)^T \mathbf{V}(t_i)^{-1} \mathbf{e}(t_i) \tag{2.8}$$

$$\Delta \leftarrow \Delta + ln|\mathbf{V}(t_i)| \tag{2.9}$$

where $\leftarrow$ indicates that the left side is replaced by the right side as in a computer program.

5. Update the estimate of the state vector and its covariance matrix by letting

$$\mathbf{A}(t_i) = \mathbf{H}(t_i)\mathbf{P}(t_i|t_{i-1}) \tag{2.10}$$

and then setting

$$\mathbf{s}(t_i|t_i) = \mathbf{s}(t_i|t_{i-1}) + \mathbf{A}(t_i)^T \mathbf{V}(t_i)^{-1} \mathbf{e}(t_i) \tag{2.11}$$

and

$$\mathbf{P}(t_i|t_i) = \mathbf{P}(t_i|t_{i-1}) - \mathbf{A}(t_i)^T \mathbf{V}(t_i)^{-1} \mathbf{A}(t_i). \tag{2.12}$$

Now return to step 1 and iterate through the data. If n is the total number of observations, $-2ln$ likelihood is calculated as:

$$l = nln(2\pi) + \Delta + RSS. \tag{2.13}$$

2.3 Using the Kalman recursion to calculate the likelihood function of an AR(1) process

A simple example to illustrate the use of the Kalman recursion to calculate likelihoods, is to use it to calculate the likelihood of an AR(1) process. A zero mean AR(1) process $\{\mathbf{y}(t_i)\}$ is already written in state-space from. The state vector is scalar, $s(t_i)$, and the measurement equation can be written as

$$y(t_i) = s(t_i), \tag{2.14}$$

so the matrix $H(t_i) = 1$ and there is no observational error. The transition equation can be written as

$$s(t_i) = \phi s(t_{i-1}) + \Psi u(t_i), \tag{2.15}$$

so the state transition matrix is now the scalar autoregression coefficient ϕ . The initial conditions for the Kalman recursion are given by the mean and covariance matrix of the unconditional distribution of the state vector. In the AR(1) model, the marginal mean is zero so the initial state $s(0|0) = 0$ and the process variance is $P(0|0) = \frac{\Psi^2}{1-\phi^2}$.

Generic steps of the recursion for an AR(1) process

1. Calculate a one step prediction for the state,

$$s(t_i|t_{i-1}) = \phi s(t_{i-1}|t_{i-1}) \tag{2.16}$$

and calculate the covariance matrix of the prediction,

$$P(t_i|t_{i-1}) = \phi^2 P(t_{i-1}|t_{i-1}) + \Psi^2. \tag{2.17}$$

2. Calculate the innovation using the measurement equation $\hat{y}(t_i) = s(t_i|t_{i-1})$,

$$e(t_i) = y(t_i) - \hat{y}(t_i) = y(t_i) - s(t_i|t_{i-1}) = y(t_i) - \phi s(t_{i-1}|t_{i-1}). \tag{2.18}$$

3. Calculate the covariance matrix of the innovation $e(t_i)$

$$V(t_i) = P(t_i) = \phi^2 P(t_{i-1}) + \Psi^2. \tag{2.19}$$

4. Accumulate the quantities needed to calculate the log-likelihood at the end of the recursion,

$$RSS \leftarrow RSS + \frac{(y(t_i) - \phi y(t_{i-1}))^2}{\phi^2 P(t_{i-1}) + \Psi^2} \tag{2.20}$$

$$\Delta \leftarrow \Delta + log|\phi^2 P(t_{i-1}) + \Psi^2|. \tag{2.21}$$

5. Update the estimate of the state vector and its covariance matrix,

$$s(t_i|t_i) = s(t_i|t_{i-1}) + \frac{P(t_i|t_{i-1})}{V(t_i)}(y(t_i) - \phi s(t_{i-1}|t_{i-1})) \tag{2.22}$$

$$P(t_i|t_i) = P(t_i|t_{i-1}) - \frac{P(t_i|t_{i-1})^2}{V(t_i)}. \tag{2.23}$$

now return to step 1 and iterate through the data.

To better illustrate this, the first and second iteration are written below and then the pattern is followed to obtain the likelihood.

First iteration	**Second iteration**
1. $s(1\|0) = 0$	1. $s(2\|1) = \phi y_1$
$P(1\|0) = \phi^2 \frac{\Psi^2}{1-\phi^2} + \Psi^2$	$P(2\|1) = \Psi^2$
$P(1\|0) = \frac{\Psi^2}{1-\phi^2}$	
2. $\hat{y}_1 = 0$	2. $\hat{y}_2 = \phi y_1$
$e_1 = y_1$	$e_2 = y_2 - \phi y_1$
3. $V_1 = P_1 = \frac{\Psi^2}{1-\phi^2}$	3. $V_2 = \Psi^2$
4. $RSS \leftarrow \frac{y_1^2(1-\phi^2)}{\Psi^2}$	4. $RSS \leftarrow RSS + \frac{(y_2-\phi y_1)^2}{\Psi^2}$
$\Delta \leftarrow log(\frac{\Psi^2}{1-\phi^2})$	$\Delta \leftarrow \Delta + log(\Psi^2)$
5. $s(1\|1) = y_1$	5. $s(2\|2) = y_2$
$P(1\|1) = 0$	$P(2\|2) = 0$

After applying the Kalman filter recursion to the n observations we obtain that,

$$RSS = \frac{y_1^2(1-\phi^2)}{\Psi^2} + \sum_{t=2}^{n} \frac{(y_t - \phi y_{t-1})^2}{\Psi^2} \tag{2.24}$$

and,

$$\Delta = log\left(\frac{\Psi^2}{1-\phi^2}\right) + (n-1)log\Psi^2. \tag{2.25}$$

Therefore we obtain the AR(1) process -2ln likelihood as,

$$\begin{aligned} l(\Psi^2, \phi) &= n\log(2\pi) + log\left(\frac{\Psi^2}{1-\phi^2}\right) + (n-1)log\Psi^2 + \frac{y_1^2(1-\phi^2)}{\Psi^2} + \sum_{t=2}^{n}\frac{(y_t - \phi y_{t-1})^2}{\Psi^2} \\ &= n\log(2\pi) + n\log(\Psi^2) - \log(1-\phi^2) + \frac{1}{\Psi^2}[y_1^2(1-\phi^2) + \sum_{t=2}^{n}(y_t - \phi y_{t-1})^2]. \end{aligned}$$

Let

$$F(\phi) = y_1^2(1-\phi^2) + \sum_{t=2}^{n}(y_t - \phi y_{t-1})^2$$

then $l(\Psi^2, \phi)$ is given by

$$n\log(2\pi) + n\log(\Psi^2) - \log(1-\phi^2) + \frac{1}{\Psi^2}F(\phi).$$

This can be obtained directly as a check. Now, to concentrate Ψ^2 out of the likelihood we note that

$$\frac{\partial}{\partial\Psi^2}l(\Psi^2, \phi) = \frac{n}{\Psi^2} - \frac{F(\phi)}{\Psi^4}$$

$$\frac{\partial}{\partial\Psi^2}l(\Psi^2, \phi) = 0 \iff \hat{\Psi}^2 = \frac{F(\phi)}{n}$$

and,

$$l(\hat{\Psi}^2, \phi) = n\log(2\pi) + n\log\left(\frac{F(\phi)}{n}\right) - \log(1-\phi^2) + n.$$

Now the profile likelihood $l(\hat{\Psi}^2, \phi)$ can be minimized over ϕ to get $\hat{\phi}$ and then calculate $\hat{\Psi}^2$ using

$$\hat{\Psi}^2 = \frac{F(\hat{\phi})}{n}. \tag{2.26}$$

Suppose that, for the purpose of running the Kalman filter, Ψ^2 is replaced by 1. Therefore, the iterations on page 10 can be run with $\mathbf{P}(0|0) = \frac{1}{1-\phi^2}$. After applying the Kalman filter recursion to the n observations, it is found that

$$RSS^* = y_1^2(1-\phi^2) + \sum_{t=2}^{n}(y_t - \phi y_{t-1})^2 \tag{2.27}$$

and

$$\Delta^* = -\log(1-\phi^2). \tag{2.28}$$

The log-likelihood comes out to be (up to an additive constant)

$$l^*(\phi) = \Delta^* + n\log(RSS^*).$$

Also note that $RSS^* = F(\phi)$ hence,

$$\begin{aligned} l^*(\phi) &= -\log(1-\phi^2) + n\log(F(\phi)) + constant \\ l^* + constant &= l(\hat{\Psi}^2, \phi). \end{aligned}$$

Now $l^*(\phi)$ is minimized to estimate $\hat{\phi}$ and then

$$\hat{\Psi}^2 = \frac{RSS^*}{n} = \frac{F(\hat{\phi})}{n},$$

which is the same as equation (2.26). Hence by replacing Ψ^2 by 1 in the Kalman filter, at the end of the iterations the profile likelihood $l(\hat{\Psi}^2, \phi)$ is constructed directly.

Another possibility to obtain directly the profile likelihood $l(\hat{\Psi}^2, \phi)$ is to start the recursion with $\mathbf{P}(0|0) = 1$ and therefore replace (for the purpose of running the filter only) Ψ^2 by $1 - \phi$. At the end of this run, it is found that

$$RSS^{**} = y_1^2 + \sum_{t=2}^{n} \frac{(y_t - \phi y_{t-1})^2}{1 - \phi^2} = \frac{F(\phi)}{1 - \phi^2} \tag{2.29}$$

and

$$\Delta^{**} = (n-1)\log(1 - \phi^2). \tag{2.30}$$

Then

$$\begin{aligned} l^{**}(\phi) &= \Delta^{**} + n\log(RSS^{**}) \\ &= (n-1)\log(1-\phi^2) + n\log\left(\frac{F(\phi)}{1-\phi^2}\right) \\ &= (n-1)\log(1-\phi^2) + n\log(F(\phi)) - n\log(1-\phi^2) \\ &= n\log(F(\phi)) - \log(1-\phi^2) \\ &= l^*(\phi), \end{aligned}$$

also,

$$l^{**}(\phi) + constant = l(\hat{\Psi}^2, \phi),$$

and

$$\hat{\Psi}^2 = (1 - \hat{\phi}^2)\frac{RSS^{**}}{n}.$$

Replacing Ψ^2 by 1 or by $1-\phi^2$ to run the Kalman filter are methods to obtain the profile likelihood $l(\hat{\Psi}^2, \phi)$ directly without first having to minimize the likelihood $l(\Psi^2, \phi)$ with respect to Ψ^2.

2.4 Unequally spaced data

When observations can be taken at arbitrary time points, there must be an underlying continuous time process (Jones 1981, 1985; Diggle 1988, 1990; Chi and Reinsel 1989). The equally spaced AR(1) model is a difference equation driven by a sequence of independent random variables called white noise (Box and Jenkins 1976). A model for a continuous time process is a differential equation, also driven by white noise. Continuous time white noise exists only in the sense that its integral is a continuous time random walk $\{w(t)\}$ often referred to as a Brownian motion or a Wiener process.

The model for a continuous time AR(1) process, denoted CAR(1) is

$$dy(t) + \alpha y(t)dt = \sigma dw(t). \tag{2.31}$$

Even though the derivative of $w(t)$ does not exist, when the differential equation (2.31) is solved by integration, a proper solution is obtained. The Wiener process $w(t)$ is assumed to have unit variance per unit time, and the constant σ in front of $dw(t)$ scales the input so it has variance $\sigma^2 dt$.

If the solution to the differential equation (2.31) with the random input $dw(t)$ replaced by zero is considered and it is integrated from time t_1 to time t_2, the solution is a prediction

$$y(t_2) = exp\{-\alpha(t_2 - t_1)\}y(t_1). \tag{2.32}$$

Therefore the solution is in the same form as a discrete time AR(1) process with autoregresive coefficient,

$$\phi(t_2 - t_1) = exp\{-\alpha(t_2 - t_1)\}, \tag{2.33}$$

which depends on the time interval.

The condition for stationarity of a CAR(1) process is $\alpha > 0$ which implies

$$0 < \phi(t_2 - t_1) \leq 1, \tag{2.34}$$

and $\phi(t_2 - t_1)$ can take the value 1 only when the time interval is zero. (This is different from the usual discrete time AR(1) process.)

The solution for the differential equation (2.31) from time zero to time t is

$$y(t) = e^{-\alpha t}y(0) + \sigma \int_0^t e^{-\alpha(t-r)}dw(r). \tag{2.35}$$

While integrating the process from t_1 to t_2, small random inputs, $dw(t)$, enter at each time. As the influence of the initial condition dies away exponentially with time, so does the effect of the random inputs. At time t_2, the influence of the random input at some earlier time is down weighted by the factor $exp\{-\alpha(t_2 - t_1)\}$. The result can be integrated over the time interval to determine the properties of the random input over a finite interval

$$u(t_2 - t_1) = \sigma \int_{t_1}^{t_2} exp\{-\alpha(t_2 - t)\}dw(t). \tag{2.36}$$

From the properties of a Wiener process, the infinitesimally small increments, $dw(t)$, are independent with zero mean and variance dt, so $E[u(t_2 - t_1)] = 0$. To calculate the

variance of $u(t_2 - t_1)$, the variances of small increments can be integrated over the time interval remembering that the variance of a constant multiplied by a random variable is the constant squared multiplied by the variance of the random variable. Then

$$var[u(t_2 - t_1)] = \sigma^2 \int_{t_1}^{t_2} exp\{-2\alpha(t_2 - t)\}dt \tag{2.37}$$

$$= \frac{\sigma^2}{2\alpha}[1 - exp\{-2\alpha(t_2 - t_1)\}] \tag{2.38}$$

The assumption of a Gaussian distribution for the integral of $u(t)$ implies that $u(t)$ must also have a Gaussian distribution. This leads to the usual prediction error decomposition form of the likelihood function. The exact likelihood function is obtained by including the unconditional (marginal) distribution of y_1. If in equation (2.35) time $t = 0$ is replaced by $t = -\infty$ we obtain

$$y(t) = e^{-\alpha t}y(-\infty) + \sigma \int_{-\infty}^{t} e^{-\alpha(t-r)}dw(r). \tag{2.39}$$

Now if $y(-\infty)$ is set equal to the mean zero,

$$y(t) = \sigma \int_{-\infty}^{t} e^{-\alpha(t-r)}dw(r), \tag{2.40}$$

but since before $y(0)$ was set equal to $y(-\infty)$ then $y_1 = y(0)$ which is the observation at $t = 0$, and we obtain

$$y_1 = y(0) = \sigma \int_{-\infty}^{0} e^{-\alpha(0-r)}dw(r) = \sigma \int_{-\infty}^{0} e^{\alpha r}dw(r). \tag{2.41}$$

Therefore, the unconditional (marginal) variance of y_1 is

$$var(y_1) = \sigma^2 \int_{-\infty}^{0} e^{2\alpha r} dr = \frac{\sigma^2}{2\alpha}. \tag{2.42}$$

2.5 Using the Kalman recursion to calculate the likelihood function of a CAR(1) process

Following section (2.4), it is possible to write a CAR(1) process in state-space form, where the measurement equation can be written as

$$y(t_i) = s(t_i), \tag{2.43}$$

and the transition equation can be written as

$$s(t_i) = \phi(t_i - t_{i-1})s(t_{i-1}) + \Psi(t_i - t_{i-1})u(t_i) \tag{2.44}$$

where $\phi(t_i - t_{i-1}) = exp\{-\alpha(t_i - t_{i-1})\}$ and

$$var[u(t_i - t_{i-1})] = \Psi(t_i - t_{i-1})^2 = \frac{\sigma^2}{2\alpha}[1 - exp\{-2\alpha(t_i - t_{i-1})\}]. \tag{2.45}$$

To apply the Kalman recursion, it is necessary to set the initial values. Since in continuous processes there is no natural time to be assigned 0, the initial values are called $s(t_1|0)$ and $P(t_1|0)$. As mentioned in section (2.3), the initial conditions for the Kalman recursion are given by the mean and variance of the unconditional distribution of the state vector. In the CAR(1) model, its mean is zero so the initial state $s(t_1|0) = 0$ and the initial process variance is $P(t_1|0) = \frac{\sigma^2}{2\alpha}$. To better illustrate how the Kalman

filter works, two iterations are produced, and then the likelihood of a CAR(1) process is obtained.

First iteration	**Second iteration**
1. $s(t_1\|0) = 0$	1. $s(t_2\|t_1) = \exp\{-\alpha(t_2 - t1)\}y_1$
$P(t_1\|0) = \frac{\sigma^2}{2\alpha}$	$P(t_2 - t_1) = \frac{\sigma^2}{2\alpha}[1 - exp\{-2\alpha(t_2 - t_1)\}]$
2. $\hat{y}_1 = 0$	2. $\hat{y}_2 = \phi(t_2 - t_1)y_1$
$e_1 = y_1$	$e_2 = y_2 - \phi(t_2 - t_1)y_1$
3. $V_1 = \frac{\sigma^2}{2\alpha}$	3. $V_2 = \frac{\sigma^2}{2\alpha}\{1 - \phi(t_2 - t_1)^2\}$
4. $RSS \leftarrow \frac{2\alpha y_1^2}{\sigma^2}$	4. $RSS \leftarrow RSS + \frac{2\alpha[y_2 - \phi(t_2 - t_1)y_1]^2}{\sigma^2\{1 - \phi(t_2 - t_1)^2\}}$
$\Delta \leftarrow log(\frac{\sigma^2}{2\alpha})$	$\Delta \leftarrow \Delta + log\|\frac{\sigma^2}{2\alpha}\{1 - \phi(t_2 - t_1)^2\}\|$
5. $s(t_1\|t_1) = y_1$	5. $s(t_2\|t_2) = y_2$
$P(t_1\|t_1) = 0$	$P(t_2\|t_2) = 0$

After applying the Kalman recursion to the n observations we obtain

$$RSS = \frac{2\alpha}{\sigma^2}y_1^2 + \frac{2\alpha}{\sigma^2}\sum_{j=2}^{n}\frac{(y_j - \phi(t_j - t_{j-1})y_{j-1})^2}{\{1 - \phi(t_j - t_{j-1})^2\}} \tag{2.46}$$

and,

$$\Delta = log|\frac{\sigma^2}{2\alpha}| + \sum_{j=2}^{n} log|\frac{\sigma^2}{2\alpha}\{1 - \phi(t_j - t_{j-1})^2\}|. \tag{2.47}$$

Therefore we obtain the -2ln likelihood of the CAR(1) process as,

$$l = nlog(2\pi)+log|\frac{\sigma^2}{2\alpha}|+\sum_{j=2}^{n} log|\frac{\sigma^2}{2\alpha}\{1-\phi(t_j-t_{j-1})^2\}|+\frac{2\alpha}{\sigma^2}\left(y_1^2+\sum_{j=2}^{n}\frac{(y_j-\phi(t_j-t_{j-1})y_{j-1})^2}{\{1-\phi(t_j-t_{j-1})^2\}}\right). \tag{2.48}$$

This can be confirmed by direct calculation.

Chapter 3

The linear regression model

3.1 Calculating the likelihood using the Cholesky decomposition

Consider the general linear model

$$y_t = \mathbf{x_t^T}\beta + \varepsilon_t, \tag{3.1}$$

with $t = 1, \ldots, n$.

This model can be written in matrix form as

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon, \tag{3.2}$$

where $\mathbf{y}$ is a n-vector and $\mathbf{X}$ is a $n \times n$ matrix. The disturbances ε are assumed to have mean zero and variance $\mathbf{\Theta}$ and, in general, $\mathbf{\Theta}$ is non diagonal as the disturbances are serially correlated with an autoregression structure and heteroscedastic.

Since $\mathbf{\Theta}$ is positive definite, then exists a positive definite lower triangular matrix $\mathbf{L}$, with ones on the leading diagonal, and a positive definite diagonal matrix, $\mathbf{D}$, such that

$$\mathbf{\Theta}^{-1} = \mathbf{L}^{\mathbf{T}}\mathbf{D}^{-1}\mathbf{L}. \tag{3.3}$$

Pre-multiplying equation (3.2) through by $\mathbf{L}$ and defining $\mathbf{y}^*$, $\mathbf{X}^*$ and ε^* as $\mathbf{Ly}$, $\mathbf{LX}$ and $\mathbf{L}\varepsilon$ respectively gives the heteroscedastic regression model

$$\mathbf{y}^* = \mathbf{X}^*\beta + \varepsilon^*, \tag{3.4}$$

with $var(\varepsilon^*) = \mathbf{L\Theta L}^T = \mathbf{D}$ or,

$$y_t^* = \mathbf{x_t^*}\beta + \varepsilon_t^*, \quad t = 1, \ldots, n, \tag{3.5}$$

where $var(\varepsilon_t^*) = d_t$ and $d_t > 0$ is the t^{th} diagonal element of $\mathbf{D}$.

Now consider the Maximum Likelihood estimator of the full set of parameters when the disturbances ε are normally distributed. Since in equation (3.2) $\mathbf{y}$ has a multivariate normal distribution with mean $\mathbf{X}\beta$ and covariance matrix $\mathbf{\Theta}$, the log-likelihood function is

$$l = -\frac{n}{2}ln(2\pi) - \frac{1}{2}ln|\Theta| - \frac{1}{2}(\mathbf{y} - \mathbf{X}\beta)^T\Theta^{-1}(\mathbf{y} - \mathbf{X}\beta). \tag{3.6}$$

However, substituting $\mathbf{\Theta}$ from equation (3.3) and noting that

$$
\begin{aligned}
ln|\mathbf{\Theta}| &= -ln|\mathbf{\Theta}^{-1}| && (3.7)\\
&= -ln|\mathbf{L}\mathbf{D}^{-\mathbf{1}}\mathbf{L}^{\mathbf{T}}| && (3.8)\\
&= ln|\mathbf{L}||\mathbf{D}||\mathbf{L}^{T}| && (3.9)\\
&= ln|\mathbf{D}| && (3.10)\\
&= \sum_{t=1}^{n} ln(d_t), && (3.11)
\end{aligned}
$$

enables the log-likelihood function to be written as

$$
l = \frac{n}{2}ln(2\pi) - \frac{1}{2}\sum_{t=1}^{n} ln(d_t) - \frac{1}{2}\sum_{t=1}^{n}\frac{(y_t^* - x_t^{*T}\beta)^2}{d_t} \qquad (3.12)
$$

Maximizing the likelihood function (3.12) with respect to β gives the Generalized Least Squares estimator

$$
\tilde{\beta} = \left(\sum_{t=1}^{n}\frac{\mathbf{x_t^*}\mathbf{x_t^{*T}}}{d_t}\right)^{-1}\sum_{t=1}^{n}\frac{\mathbf{x^*}\mathbf{y_t^*}}{d_t} \qquad (3.13)
$$

and the residual

$$
\mathbf{e_t} = \mathbf{y_t^*} - \mathbf{x_t^{*T}}\tilde{\beta}. \qquad (3.14)
$$

3.2 Calculating the likelihood using the Kalman Filter

Harvey (1989) and Jones (1993) showed that the Kalman Filter effectively performs the Cholesky decomposition (3.3). Suppose for a moment, that β is known. In this

case, the "observations" $y_t - \mathbf{x_t^T}\beta$ can be regarded as being generated by a state-space model in which the measurement equation

$$y_t - \mathbf{x_t^T}\beta = \varepsilon_t, \quad t = 1, \ldots, n, \tag{3.15}$$

is coupled with the transition equation

$$\varepsilon_t = \phi\varepsilon_{t-1} + \Psi_t u_t. \tag{3.16}$$

Applying the Kalman Filter to these "observations" gives a set of innovations $\mathbf{e_t}$. We show that $\mathbf{e_t} = \mathbf{y_t^*} - \mathbf{x_t^{*T}}\beta$. The Kalman filter is a linear operation, so the effect of applying it to $\mathbf{y} - \mathbf{X}\beta$ is to obtain $\mathbf{A}(\mathbf{y} - \mathbf{X}\beta) = \mathbf{e}$. Also, the Cholesky decomposition is unique, so there is only one $\mathbf{L}$ such that $\mathbf{L}(\mathbf{y} - \mathbf{X}\beta) = \mathbf{y}^* - \mathbf{X}^*\beta$. The innovations $\mathbf{e_t}$ are uncorrelated with mean zero and variance d_t. This is exactly the property possessed by the disturbances $\varepsilon_t^* = \mathbf{y_t^*} - \mathbf{x_t^{*T}}\beta$ in the transformed equation (3.5). We know that $\mathbf{y_t^*} - \mathbf{x_t^{*T}}\beta$ satisfies $E(\mathbf{y_t^*} - \mathbf{x_t^{*T}}\beta) = 0$, $cov(\mathbf{y_t^*} - \mathbf{x_t^{*T}}\beta; \mathbf{y_r^*} - \mathbf{x_r^{*T}}\beta) = 0$ when $t \neq r$ and $var(\mathbf{y_t^*} - \mathbf{x_t^{*T}}\beta) = d_t$. Therefore, it follows that, $\mathbf{A} = \mathbf{L}$ and hence

$$\mathbf{e_t} = \mathbf{y_t^*} - \mathbf{x_t^{*T}}\beta, \quad t = 1, \ldots, n. \tag{3.17}$$

Therefore, in matrix terms, the effect of the Kalman Filter is to produce the innovation vector $\mathbf{e} = (e_1, \ldots, e_n)^T$ by pre-multiplying the vector $\mathbf{y} - \mathbf{X}\beta$ by the $n \times n$ matrix $\mathbf{L}$ defined in equation (3.3). However, since

$$\mathbf{L}(\mathbf{y} - \mathbf{X}\beta) = \mathbf{L}\mathbf{y} - \mathbf{L}\mathbf{X}\beta \tag{3.18}$$

$$= \mathbf{y}^* - \mathbf{X}^*\beta, \tag{3.19}$$

the same Kalman Filter can be applied separately to the observations $\mathbf{y_t}$ and each of the explanatory variables in the vector $\mathbf{x_t}$. The Generalized Least Squares estimator $\tilde{\beta}$ is then computed by regressing $\mathbf{y_t^*}$ (which are the "innovations" from applying the Kalman Filter to $\mathbf{y_t}$) on the $\mathbf{x_t^*}$ (which are the "innovations" from applying the Kalman Filter to $\mathbf{x_t}$.

To better illustrate how the Kalman Filter works for a regression model a simple example is presented. Consider the model

$$\mathbf{y} = \mathbf{x}\beta + \varepsilon \tag{3.20}$$

with

$$\varepsilon_t = \phi\varepsilon_{t-1} + \Psi_t u_t, \tag{3.21}$$

where, if n is the number of observations, $\mathbf{y}$ is an n-vector of observed responses, $\mathbf{x}$ is an n-vector of observed explanatory variables, ε is an n-vector of disturbances with mean zero and variance one, and u_t is the random input to the state at time t with mean zero and covariance $\Psi^2\sigma^2$.

As it was mentioned earlier in this section, it is necessary to apply the Kalman filter to both, the responses $\mathbf{y}$ and to the explanatory variables $\mathbf{x}$. Jones (1993) and Durbin and Koopman (2001) showed that the state vector can be augmented and the measurement equation can be written as

$$(\mathbf{x}, \mathbf{y}) = (\mathbf{s_x}, \mathbf{s_y}) \tag{3.22}$$

where $(\mathbf{x}, \mathbf{y})$ is an $n \times 2$ matrix of observations, $\mathbf{s}_x$ is the unobserved state vector for x and $\mathbf{s}_y$ is the unobserved state vector for y. The rows of $[\mathbf{x}, \mathbf{y}]$ are $[x_t, y_t]$ and the rows

of $(\mathbf{s}_x, \mathbf{s}_y)$ are $(s_x(t), s_y(t))$. The transition equation is

$$\varepsilon_t = \phi\varepsilon_{t-1} + \Psi_t u_t, \tag{3.23}$$

where $\varepsilon_t = s_x(t) = s_y(t)$.

It is also needed to set the initial state vectors and their variances. In this case, $s_x(0|0) = s_y(0|0) = 0$ and $\mathbf{P}_x(0|0) = \mathbf{P}_y(0|0) = 1$. Note that from section (2.3), the variance of an AR(1) process is $\frac{\Psi^2}{1-\phi^2}$ and since this variance was set to one, $\Psi^2 = 1-\phi^2$.

In step 4 of the Kalman filter, it is necessary a 2×2 matrix to accumulate the quantities needed to calculate the log-likelihood at the end of the recursion. This matrix is called $\mathbf{M}$ and initially it is set to zero.

Thus, the Kalman filter can be run starting with

$$s_x(0|0) = s_y(0|0) = 0,$$

$$\mathbf{P}_x(0|0) = \mathbf{P}_y(0|0) = 1,$$

$$\Psi^2 = 1 - \phi^2,$$

$$\mathbf{M} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix},$$

and

$$\Delta = 0.$$

Now the first iteration is started and a one step prediction of the state and its variance at time 1 is calculated. Letting s represent either s_x or s_y and $\mathbf{P}$ represent either $\mathbf{P}_x$ or $\mathbf{P}_y$, we obtain

$$s(1|0) = \phi s(0|0)$$

$$\begin{aligned}
\mathbf{P}(1|0) &= \phi^2 \mathbf{P}(0|0) + \Psi^2 \\
&= \phi^2 + (1 - \phi^2) \\
&= 1.
\end{aligned}$$

Calculate the prediction of the next observation vector and calculate the innovation vector

$$\begin{pmatrix} \hat{x_1} & \hat{y_1} \end{pmatrix} = \begin{pmatrix} s_x(1|0) & s_y(1|0) \end{pmatrix} = \begin{pmatrix} 0 & 0 \end{pmatrix}$$

$$\mathbf{e}_1 = \begin{pmatrix} x_1 & y_1 \end{pmatrix} - \begin{pmatrix} \hat{x_1} & \hat{y_1} \end{pmatrix} = \begin{pmatrix} x_1 & y_1 \end{pmatrix}.$$

Calculate the variance of the innovation vector $\mathbf{e}_1$ as

$$\mathbf{V}_1 = P(1|0) = 1,$$

for both x and y.

Accumulate the quantities needed to calculate the log-likelihood at the end of the recursion

$$\mathbf{M} \leftarrow \mathbf{M} + e_1^T V_1^{-1} e_1$$

$$\mathbf{M} \leftarrow \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} \begin{pmatrix} x_1 & y_1 \end{pmatrix}$$

$$\mathbf{M} \leftarrow \begin{pmatrix} x_1^2 & x_1 y_1 \\ x_1 y_1 & y_1^2 \end{pmatrix}$$

$$\Delta \leftarrow \Delta + \log |\mathbf{V}_1|$$

$$\Delta \leftarrow \Delta + \log |1| = 0.$$

Update the estimate of the state and its covariance by

$$\begin{pmatrix} s_x(1|1) & s_y(1|1) \end{pmatrix} = \begin{pmatrix} s_x(1|0) & s_y(1|0) \end{pmatrix} + \begin{pmatrix} \frac{\mathbf{P}(1|0)}{\mathbf{V}_1} x_1 & \frac{\mathbf{P}(1|0)}{\mathbf{V}_1} y_1 \end{pmatrix}$$

$$= \begin{pmatrix} x_1 & y_1 \end{pmatrix}$$

and, for either x or y,

$$\mathbf{P}(1|1) = \mathbf{P}(1|0) - \frac{\mathbf{P}(1|0)^2}{\mathbf{V}_1} = 0.$$

After two iterations we obtain

$$\begin{pmatrix} s_x(2|1) & s_y(2|1) \end{pmatrix} = \begin{pmatrix} \phi x_1 & \phi y_1 \end{pmatrix},$$

$$\mathbf{P}(2|1) = \Psi^2 = 1 - \phi^2,$$

$$\mathbf{e}_2 = \begin{pmatrix} (x_2 - \phi x_1) & (y_2 - \phi y_1) \end{pmatrix}$$

$$\mathbf{M} \leftarrow \begin{pmatrix} x_1^2 + (x_2 - \phi x_1) & x_1 y_1 + (x_2 - \phi x_1)(y_2 - \phi y_1) \\ x_1 y_1 + (x_2 - \phi x_1)(y_2 - \phi y_1) & y_1^2 + (y_2 - \phi y_1) \end{pmatrix}$$

and

$$\Delta \leftarrow \log(1 - \phi^2).$$

After n iterations, where n is the number of observations we obtain

$$\mathbf{M} = \begin{pmatrix} x_1^2 + \sum_{t=2}^{n} \frac{(x_t - \phi x_{t-1})^2}{1-\phi^2} & x_1 y_1 + \sum_{t=2}^{n} \frac{(x_t - \phi x_{t-1})(y_t - \phi y_{t-1})}{1-\phi^2} \\ x_1 y_1 + \sum_{t=2}^{n} \frac{(x_t - \phi x_{t-1})(y_t - \phi y_{t-1})}{1-\phi^2} & y_1^2 + \sum_{t=2}^{n} \frac{(y_t - \phi y_{t-1})^2}{1-\phi^2} \end{pmatrix}$$

and

$$\Delta = (n-1)\log(1 - \phi^2).$$

Note that

$$\mathbf{M} = \begin{pmatrix} \mathbf{x}^{*T}\mathbf{x}^* & \mathbf{x}^{*T}\mathbf{y}^* \\ \mathbf{y}^{*T}\mathbf{x}^* & \mathbf{y}^{*T}\mathbf{y}^* \end{pmatrix}$$

where

$$\mathbf{x}^* = \mathbf{L}\mathbf{x} = \begin{pmatrix} x_1 \\ (x_2 - \phi x_1)/\sqrt{1-\phi^2} \\ \vdots \\ (x_n - \phi x_{n-1})/\sqrt{1-\phi^2} \end{pmatrix}$$

and

$$\mathbf{y}^* = \mathbf{L}\mathbf{y} = \begin{pmatrix} y_1 \\ (y_2 - \phi y_1)/\sqrt{1-\phi^2} \\ \vdots \\ (y_n - \phi y_{n-1})/\sqrt{1-\phi^2} \end{pmatrix}.$$

The solution for β of the transformed equation (3.19) is the usual least squares solution

$$\hat{\beta} = (\mathbf{x}^{*T}\mathbf{x}^*)^{-1}\mathbf{x}^{*T}\mathbf{y}^*.$$

The total sum of squares is $\mathbf{y}^*\mathbf{y}^{*T}$ and the residual sum of squares is

$$\begin{aligned} RSS &= (\mathbf{y}^* - \mathbf{X}^*\hat{\beta})^T(\mathbf{y}^* - \mathbf{X}^*\hat{\beta}) && (3.24) \\ &= \mathbf{y}^{*T}\mathbf{y}^* - \mathbf{y}^*\mathbf{x}^{*T}\hat{\beta} && (3.25) \\ &= \mathbf{y}^{*T}\mathbf{y}^* - \mathbf{y}^{*T}\mathbf{x}^*(\mathbf{x}^{*T}\mathbf{x}^*)^{-1}\mathbf{x}^{*T}\mathbf{y}^* && (3.26) \end{aligned}$$

so the estimate of Ψ^2 is

$$\hat{\Psi}^2 = \frac{1}{n}RSS.$$

Therefore the -2log-likelihood with Ψ^2 concentrated out of it, can be written as

$$\begin{aligned} l &= nln(2\pi\hat{\Psi}^2) + ln|\mathbf{V}| + n \\ &= nln(2\pi) + nln(\hat{\Psi}^2) + \Delta + n \\ &= n + n\log(2\pi) + (n-1)\log(1-\phi^2) + nln[\frac{1}{n}(\mathbf{y}^{*T}\mathbf{y}^* - \mathbf{y}^{*T}\mathbf{x}^*(\mathbf{x}^{*T}\mathbf{x}^*)^{-1}\mathbf{x}^{*T}\mathbf{y}^*] \end{aligned}$$

$$
\begin{aligned}
&= n + n\log(2\pi) + (n-1)\log(1-\phi^2) - n\log(n) + n\log\left(\mathbf{y}^{*T}\mathbf{y}^{*} - \mathbf{y}^{*T}\mathbf{x}^{*}(\mathbf{x}^{*T}\mathbf{x}^{*})^{-1}\mathbf{x}^{*T}\mathbf{y}^{*}\right) \\
&= n + n\log(2\pi) + (n-1)\log(1-\phi^2) - n\log(n) + \\
&\quad + n\log\left(y_1^2 + \sum_{t=2}^{n}\frac{(y_t-\phi y_{t-1})^2}{1-\phi^2} - \frac{\left(x_1y_1 + \sum_{t=2}^{n}\frac{(x_t-\phi x_{t-1})(y_t-\phi y_{t-1})}{1-\phi^2}\right)^2}{x_1^2 + \sum_{t=2}^{n}\frac{(x_t-\phi x_{t-1})^2}{1-\phi^2}}\right).
\end{aligned}
$$

Now the -2 log-likelihood is a function of ϕ only and by minimizing it with respect of ϕ it is possible to obtain the maximum likelihood estimator of ϕ.

It is also possible to use the substitution $\Psi^2 = 1$ in the Kalman filter. In this case the initial values will be,

$$s_x(0|0) = s_y(0|0) = 0,$$

$$\mathbf{P}_x(0|0) = \mathbf{P}_y(0|0) = \frac{1}{1-\phi^2},$$

$$\Psi^2 = 1,$$

$$\mathbf{M} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix},$$

and

$$\Delta = 0.$$

After the first iteration we obtain,

$$s_x(1|0) = s_y(1|0) = 0,$$

$$\mathbf{P}_x(1|0) = \mathbf{P}_y(1|0) = \frac{1}{1-\phi^2},$$

$$\mathbf{M} \leftarrow (1-\phi^2)\begin{pmatrix} x_1^2 & x_1y_1 \\ x_1y_1 & y_1^2 \end{pmatrix},$$

$$\Delta \leftarrow -\log(1-\phi^2),$$

$$\begin{pmatrix} s_x(1|1) & s_y(1|1) \end{pmatrix} = \begin{pmatrix} x_1 & y_1 \end{pmatrix}$$

and

$$\mathbf{P}(1|1) = 0.$$

If another iteration is run, it shows that,

$$\begin{pmatrix} s_x(2|1) & s_y(2|1) \end{pmatrix} = \begin{pmatrix} \phi x_1 & \phi y_1 \end{pmatrix},$$

$$\mathbf{P}_x(2|1) = \mathbf{P}_y(2|1) = 1,$$

$$\mathbf{M} \leftarrow \begin{pmatrix} x_1^2(1-\phi^2) + (x_2 - \phi x_1) & x_1 y_1(1-\phi^2) + (x_2 - \phi x_1)(y_2 - \phi y_1) \\ x_1 y_1(1-\phi^2) + (x_2 - \phi x_1)(y_2 - \phi y_1) & y_1^2(1-\phi^2) + (y_2 - \phi y_1) \end{pmatrix},$$

$$\Delta \leftarrow -\log(1-\phi^2),$$

$$\begin{pmatrix} s_x(2|2) & s_y(2|2) \end{pmatrix} = \begin{pmatrix} x_2 & y_2 \end{pmatrix}$$

and

$$\mathbf{P}(2|2) = 0.$$

After iterating through all the observations we obtain

$$\mathbf{M} \leftarrow \begin{pmatrix} x_1^2(1-\phi^2) + \sum_{t=2}^{n}(x_t - \phi x_{t-1})^2 & x_1 y_1(1-\phi^2) + \sum_{t=2}^{n}(x_t - \phi x_{t-1})(y_t - \phi y_{t-1}) \\ x_1 y_1(1-\phi^2) + \sum_{t=2}^{n}(x_t - \phi x_{t-1})(y_t - \phi y_{t-1}) & y_1^2(1-\phi^2) + \sum_{t=2}^{n}(y_t - \phi y_{t-1})^2 \end{pmatrix}$$

and

$$\Delta \leftarrow -\log(1-\phi^2),$$

It is very important to notice that when this substitution is used, the matrix $\mathbf{M}$ does not contain the Cholesky decomposition, but instead contains $x^* x^{*T}(1-\phi^2)$, $x^* y^{*T}(1-\phi^2)$ and $y^* y^{*T}(1-\phi^2)$. It is necessary to be very careful when using these transformation because although intuitively it might look like they should produce the same result in different models, in fact they do not.

3.3 The Laird - Ware model

Based on the work of Harville (1974, 1976, 1977), Laird and Ware proposed a very general linear mixed model for longitudinal data,

$$\mathbf{y_i} = \mathbf{X_i}\beta + \mathbf{Z_i}\gamma_\mathbf{i} + \varepsilon_\mathbf{i}, \tag{3.27}$$

where $\mathbf{y_i}$ is an $n_i \times 1$ column vector of response variable for subject i, $\mathbf{X_i}$ is an $n_i \times p$ design matrix, β is a $p \times 1$ of regression coefficients assumed to be fixed, $\mathbf{Z_i}$ is an $n_i \times q$ design matrix for the random effects and $\gamma_\mathbf{i}$, which are assumed to be normally distributed with mean zero and variance $\sigma^2\mathbf{B}$ independently distributed across subjects. The matrix $\mathbf{B}$ is an arbitrary covariance matrix. The within subject errors, $\varepsilon_\mathbf{i}$ are assumed to be normally distributed with mean zero and variance $\sigma^2\mathbf{W_i}$, where $\mathbf{W_i}$ is a covariance matrix. The $\varepsilon_\mathbf{i}$ are also independent from subject to subject and independent of $\gamma_\mathbf{i}$.

The Laird - Ware model is very general since different subjects can have different numbers of observations as well as different observation times.

Even though the Laird - Ware model is much more general, its likelihood function is similar to the likelihood for a simple linear mixed model. When there are several groups of subjects, this is incorporated into the design matrix $\mathbf{X_i}$, and then the mean vector for subject i is $\mathbf{X_i}\beta$. Since the two random components have zero means, $\gamma_\mathbf{i}$ is uncorrelated with $\varepsilon_\mathbf{i}$ and the two covariance matrices are $cov(\gamma_\mathbf{i}) = \sigma^2\mathbf{B}$ and $cov(\varepsilon_\mathbf{i}) = \sigma^2\mathbf{W_i}$, the total covariance matrix for subject i is

$$\sigma^2\mathbf{G} = \sigma^2(\mathbf{Z_i}\mathbf{B}\mathbf{Z_i^T} + \mathbf{W_i}). \tag{3.28}$$

Assuming that there are n_i observations for subject i, the -2ln likelihood is

$$l = \sum_i [n_i log(2\pi) + log|\sigma^2\mathbf{G_i}| + (\mathbf{y_i} - \mathbf{X_i}\beta)^\mathbf{T}(\sigma^\mathbf{2}\mathbf{G_i})^{-1}(\mathbf{y_i} - \mathbf{X_i}\beta)]. \tag{3.29}$$

Differentiating l with respect to σ^2 and setting the result to zero gives

$$\hat{\sigma}^2 = \frac{1}{n}\sum_i (\mathbf{y}_i - \mathbf{X}_i\beta)^T\mathbf{G}_i(\mathbf{y}_i - \mathbf{X}_i\beta). \tag{3.30}$$

Also, for given values of the $\mathbf{B}$ and $\mathbf{W}_i$ matrices, the estimate of β that minimizes l is

$$\hat{\beta} = (\sum_i \mathbf{X}_i^T\mathbf{G}_i^{-1}\mathbf{X}_i)^{-1}(\sum_i \mathbf{X}_i^T\mathbf{G}_i^{-1}\mathbf{y_i}), \tag{3.31}$$

the generalized least squares estimator.

3.4 The Laird-Ware model in state-space form

The Laird-Ware model can be put in state-space form in a similar way as a regression model, but it is necessary to include the random effects. Duncan and Horn (1972) showed that mixed models can be put in state-space form by including the random effects in the state vector.

Consider the model in equation (3.27) and the within subject errors ε_i are assumed to have an AR(1) structure. It is possible to write this model in state-space form with the observation equation

$$(\mathbf{x}_i(t_{ij})^T, y_i(t_{ij})) = \mathbf{h}\mathbf{S}_i(t_{ij}). \tag{3.32}$$

As in the regression case, it is necessary to apply the Kalman Filter to both the responses y and to the explanatory variables x, so the observation vector is the p-row vector $\mathbf{x}_i(t_{ij})$ transposed augmented by $y_i(t_{ij})$, $\mathbf{h}$ is a $(q+1)$-row vector equal to

$$\mathbf{h} = \begin{pmatrix} 1 & \mathbf{z}_i(t_{ij}) \end{pmatrix}.$$

The state $\mathbf{S}_i(t_{ij})$ is now a $(q+1) \times (p+1)$ matrix,

$$\mathbf{S}_i(t_{ij}) = \begin{pmatrix} s_{x_1}(t_{ij}) & \dots & s_{x_p}(t_{ij}) & s_y(t_{ij}) \end{pmatrix} = \begin{pmatrix} \varepsilon_i(t_{ij}) & \dots & \varepsilon_i(t_{ij}) \\ \gamma_i(t_{ij}) & \dots & \gamma_i(t_{ij}) \end{pmatrix}.$$

The transition equation can be written as

$$\begin{pmatrix} \varepsilon_i(t_{ij}) \\ \gamma_i(t_{ij}) \end{pmatrix} = \begin{pmatrix} \phi & 0 \\ 0 & I \end{pmatrix} \begin{pmatrix} \varepsilon_i(t_{i,j-1}) \\ \gamma_i(t_{i,j-1}) \end{pmatrix} + \begin{pmatrix} \Psi_i(t_{ij})u(t_{ij}) \\ 0 \end{pmatrix},$$

where

$$\begin{pmatrix} \varepsilon_i(t_{ij}) \\ \gamma_i(t_{ij}) \end{pmatrix} = s_{x_1}(t_{ij}) = \ldots = s_{x_p}(t_{ij}) = s_y(t_{ij}).$$

Now it is necessary to set the initial state vectors and their variances and initialize the matrix M and the scalar Δ. The initial values of the Kalman Filter in this case are

$$s_{x_1}(0|0) = \ldots = s_{x_p}(0|0) = s_y(0|0) = 0,$$

$$\mathbf{P}_{x_1}(0|0) = \ldots = \mathbf{P}_{x_p}(0|0) = \mathbf{P}_y(0|0) = \begin{pmatrix} 1 & 0 \\ 0 & \mathbf{B} \end{pmatrix},$$

$$\Psi^2 = 1 - \phi^2,$$

$$\mathbf{M} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$$

and

$$\Delta = 0.$$

Now the Kalman filter can be run as in Section 3.2 and at the end the -2log-likelihood of the Laird-Ware model as a function of ϕ and the parameters in the matrix $\mathbf{B}$ is obtained.

It is very important to note that in the Laird-Ware model it is necessary to re-initialize the state and its variance after finishing with each subject i, while $\mathbf{M}$ and Δ keep accumulating throughout all the subjects.

Chapter 4

Using the Kalman filter to obtain the maximum likelihood estimators of the parameters

4.1 The likelihood function

Suppose there are T sets of observations $y_1, \ldots, y_T$ which are independent and identically distributed. Their joint density function is therefore given by

$$L(y;\Omega) = \Pi_{t=1}^{T} p(y_t) \tag{4.1}$$

where $p(y_t)$ is the joint probability density function of the t^{th} set of observations.

When the observations are not independent, the joint density function can be written as

$$L(y;\Omega) = \Pi_{t=1}^{T} p(y_t|Y_{t-1}) \tag{4.2}$$

where $p(y_t|Y_{t-1})$ denotes the distribution of y_t conditional on the information set at time t-1, that is $Y_{t-1} = \{y_{t-1}, y_{t-2}, \ldots, y_1\}$.

If the disturbances and initial state vector in the measurement equation have proper multivariate normal distributions, the distribution of y_t, conditional on Y_{t-1}, is itself normal. The mean and covariance matrix of this conditional distribution are given directly by the Kalman filter.

Harvey (1989) showed that from the derivation of the Kalman filter it is possible to see that, conditional on Y_{t-1}, $s(t)$ is normally distributed with mean $s(t|t-1)$ and covariance matrix $\mathbf{P}(t|t-1)$. If the measurement equation is written as

$$y_t = H_t s(t|t-1) + H_t(s(t) - s(t|t-1)) + \varepsilon_t$$

it can be seen that the conditional distribution of y_t is normal with mean

$$E_{t-1}(y_t) = \hat{y}(t|t-1) = H_t s(t|t-1)$$

and a covariance matrix V_t. Therefore, for a Gaussian model, the log-likelihood function of (4.2) can be written as

$$logL = -\frac{NT}{2} log 2\pi - \frac{1}{2}\sum_{t=1}^{t} log|V_t| - \frac{1}{2}\sum_{t=1}^{t} e_t^T V_t^{-1} e_t \tag{4.3}$$

were e_t is the vector of innovations.

Since the conditional mean $\hat{y}(t|t-1)$ is also the minimum mean square estimate of y_t, the innovations vector e_t can be interpreted as a vector of prediction errors. Hence (4.3) is sometimes known as the prediction error decomposition form of the likelihood.

To find the maximum likelihood estimators of the unknown parameters Ω it is necessary to maximize the likelihood function with respect to them. This is usually carried out by some kind of numerical optimization procedure like the Newton-Raphson method or Fisher's method of scoring.

4.2 Newton-Raphson method

Approximating a function by a quadratic forms the basis of relatively efficient computational schemes. A Taylor series expansion of the -2log-likelihood function $l(\Omega)$ around the minimum $\hat{\Omega}$ gives

$$l(\Omega) \approx l(\Omega) + (\Omega - \hat{\Omega})l'(\Omega) + \frac{1}{2}(\Omega - \hat{\Omega})^2 l''(\Omega). \tag{4.4}$$

Differentiating (4.4) with respect Ω gives

$$l'(\Omega) \approx l'(\hat{\Omega}) + (\Omega - \hat{\Omega})l''(\hat{\Omega}) \tag{4.5}$$

but, since $l'(\hat{\Omega}) = 0$, (4.6) can be rearrange as

$$\hat{\Omega} \approx \Omega - \frac{l'(\Omega)}{l''(\hat{\Omega})}$$

This suggests the iterative scheme

$$\Omega^{**} = \Omega^{*} - [l''(\hat{\Omega})]^{-1} l'(\Omega^{*})$$

where the revised estimate Ω^{**}, is expected to be closer to the minimum than the initial estimate Ω^{*}.

The hessian needs to be evaluated at $\hat{\Omega}$, which is unknown. Newton-Raphson solved this by evaluating the Hessian at the current estimate Ω^{*}, on the grounds that this will give an acceptable approximation to $l''(\hat{\Omega})$ if Ω^{*} is reasonable close to $\hat{\Omega}$. Therefore the iterative scheme becomes

$$\Omega^{(m)} = \Omega^{(m-1)} - [l''(\Omega^{(m-1)})]^{-1} l'(\Omega^{(m-1)}).$$

The Newton-Raphson method converges to the minimum only if $l''(\Omega)$ is positive definite.

4.3 Fisher's method of scoring

Sometimes, when maximizing the log-likelihood function l, it is easier to work with the expectation of the matrix of second derivatives rather than with the second derivatives themselves. This expectation, multiplied by minus one, gives the information matrix

$$I(\Omega) = -E[l''(\Omega)]$$

The Fisher's method of scoring uses the information matrix instead of the second derivatives matrix. Therefore, the recursive procedure becomes

$$\Omega^{(m)} = \Omega^{(m-1)} - [I(\Omega^{(m-1)})]^{-1} l'(\Omega^{(m-1)}).$$

This method is likely to have slower rate of convergence than the Newton-Raphson method because the information matrix is an approximation of the Hessian. However, in many cases, the information matrix has a simple form and is easier to compute. Also, if the model is identifiable, the information matrix is always positive definite.

4.4 Using the Kalman filter to obtain the derivatives and information matrix of the log-likelihood function

Chapters 2 and 3 showed how to use the Kalman filter to calculate the log-likelihood function of models that can be put in state-space form. However, usually the aim of obtaining the log-likelihood function of a model is to estimate the unknown parameters by using Maximum Likelihood estimation. To do this it is necessary to calculate the first and second derivatives of the log-likelihood function. Harvey (1989) showed that it is possible to calculate these derivatives using the Kalman filter.

The prediction error decomposition likelihood function (4.3) has the property that it gives an information matrix which depends on first derivatives only. These derivatives can be obtained either numerically or analytically when the model is in state-space form.

The score vector and information matrix for (4.3) can be obtained as follows. Write

the log-likelihood function in the form

$$logL = \sum_{t=1}^{T} l_t$$

where l_t is the logarithm of $p(y_t|Y_{t-1})$, ie.

$$l_t = -\frac{1}{2}log(2\pi) - \frac{1}{2}log|V_t| - \frac{1}{2}e_t^T V_t^{-1} e_t.$$

Now, for any symmetric matrix $\mathbf{A}$, the derivatives of its determinant and the inverse with respect to a variable, z, are

$$\frac{\partial|\mathbf{A}|}{\partial z} = |\mathbf{A}|tr\left(\mathbf{A}^{-1}\frac{\partial \mathbf{A}}{\partial z}\right)$$

and

$$\frac{\partial \mathbf{A}^{-1}}{\partial z} = -\mathbf{A}^{-1}\frac{\partial \mathbf{A}}{\partial z}\mathbf{A}^{-1}.$$

Differentiating l_t with respect to the i^{th} element of Ω, gives

$$-\frac{1}{2}tr\left(\mathbf{V}_t^{-1}\frac{\partial \mathbf{V}_t}{\partial \Omega_i}\right) - \frac{1}{2}\left(\frac{\partial \mathbf{e}_t^T}{\partial \Omega_i}\mathbf{V}_t^{-1}\mathbf{e}_t - \mathbf{e}_t^T\mathbf{V}_t^{-1}\frac{\partial \mathbf{V}_t}{\partial \Omega_i}\mathbf{V}_t^{-1}\mathbf{e}_t + \mathbf{e}_t^T\mathbf{V}_t^{-1}\frac{\partial \mathbf{e}_t}{\partial \Omega_i}\right).$$

Taking the trace of the last term allows this expression to be rewritten as

$$\frac{\partial l_t}{\partial \Omega_i} = -\frac{1}{2}tr\left(\left(\mathbf{V}_t^{-1}\frac{\partial \mathbf{V}_t}{\partial \Omega_i}\right)\left(\mathbf{I} - \mathbf{V}_t^{-1}\mathbf{e}_t\mathbf{e}_t^T\right)\right) - \left(\frac{\partial \mathbf{e}_t}{\partial \Omega_i}\right)^T \mathbf{V}_t^{-1}\mathbf{e_t} \qquad (4.6)$$

Differentiating (4.6) with respect to the j^{th} element of Ω gives

$$
\begin{aligned}
\frac{\partial^2 l_t}{\partial\Omega_i\partial\Omega_J} &= -\frac{1}{2}tr\left(\frac{\partial \mathbf{V}_t^{-1}\frac{\partial \mathbf{V}_t}{\partial\Omega_i}}{\partial\Omega_j}\right)\left(I-\mathbf{V}_t^{-1}\mathbf{e}_t\mathbf{e}_t^T\right) \\
&\quad -\frac{1}{2}tr\left(\mathbf{V}_t^{-1}\frac{\partial \mathbf{V}_t}{\partial\Omega_i}\mathbf{V}_t^{-1}\frac{\partial \mathbf{V}_t}{\partial\Omega_j}\mathbf{V}_t^{-1}\mathbf{e}_t\mathbf{e}_t^T\right) \\
&\quad +\frac{1}{2}tr\left(\mathbf{V}_t\frac{\partial \mathbf{V}_t}{\partial\Omega_i}\mathbf{V}_t^{-1}\left(\frac{\partial \mathbf{e}_t}{\partial\Omega_i}\mathbf{e}_t^T+\mathbf{e}_t\frac{\partial \mathbf{e}_t^T}{\partial\Omega_j}\right)\right) \\
&\quad -\frac{\partial^2\mathbf{e}_t^T}{\partial\Omega_i\partial\Omega_J}\mathbf{V}_t^{-1}\mathbf{e}_t-\frac{\partial\mathbf{e}_t^T}{\partial\Omega_i}\frac{\partial\mathbf{V}_t^{-1}}{\partial\Omega_j}\mathbf{e}_t-\frac{\partial\mathbf{e}_t^T}{\partial\Omega_i}\mathbf{V}_t^{-1}\frac{\partial\mathbf{e}_t^T}{\partial\Omega_j}. \qquad (4.7)
\end{aligned}
$$

The ij^{th} element of the information matrix is by definition

$$
-E\left(\frac{\partial^2 logL}{\partial\Omega_i\partial\Omega_J}\right)=-E\left(\sum_{t=1}^{T}\frac{\partial^2 l_t}{\partial\Omega_i\partial\Omega_J}\right),
$$

but its evaluation in the present context is simplified by noticing that

$$
E(l_t)=E\left(E_{t-1}(l_t)\right).
$$

Taking the expectation of the terms in (4.7) conditional on the information at time $t-1$, the only random variables are the elements of the innovation vector e_t, and their first and second derivatives. But the derivatives are fixed with respect to the expectation operator at time $t-1$. This is because

$$
\mathbf{e}_t=y_t-E_{t-1}(y_t)
$$

and so,

$$
\frac{\partial\mathbf{e}_t}{\partial\Omega_i}=-\frac{\partial}{\partial\Omega_i}E_{t-1}(y_t).
$$

The conditional expectation of e_t is zero and therefore,

$$E_{t-1}\left(\frac{\partial \mathbf{e}_t^T}{\partial \Omega_i}\mathbf{e}_t\right) = \frac{\partial \mathbf{e}_t^T}{\partial \Omega_i}E_{t-1}(\mathbf{e}_t) = 0.$$

A similar result holds for terms involving e_t and its second derivatives. Therefore, the third, fourth and fifth terms of (4.7) disappear. Also, the first term disappears because the conditional expectation of $e_t e_t^T$ is V_t, and for the same reason the second term simplifies. Therefore, the expression for the ij^{th} element of the information matrix is

$$I_{ij}(\Omega) = \frac{1}{2}\sum_t \left(tr\left(\mathbf{V}_t^{-1}\frac{\partial \mathbf{V}_t}{\partial \Omega_i}\mathbf{V}_t^{-1}\frac{\partial \mathbf{V}_t}{\partial \Omega_j}\right)\right) + E\left(\sum_t \left(\frac{\partial \mathbf{e}_t}{\partial \Omega_i}\right)^T \mathbf{V}_t^{-1}\frac{\partial \mathbf{e}_t}{\partial \Omega_j}\right) \qquad i,j = 1,\ldots,n. \tag{4.8}$$

Dropping the expectation operator from the second term gives an expression which is asymptotically equivalent to (4.8) and which may be easier to evaluate.

Since the i^{th} element in the score vector for (4.3) is (4.6), evaluating the score vector requires the evaluation of the N×N matrices of derivatives $\partial V_t/\partial \Omega_i$ and the N×1 vectors of derivatives $\partial e_t/\partial \Omega_i$ for $t = 1,\ldots,T$ and $i = 1,\ldots,n$. These same derivatives may be used to compute the information matrix using (4.8).

The derivatives of V_t and e_t for the state-space form, may be evaluated numerically or analytically. Computing them numerically requires n additional passes of the Kalman filter. Let δ_i for $i = 1,\ldots,n$, be a small amount added to Ω_i, the Kalman filter is run with this new value but with all the other elements in Ω remaining at their original values. This gives a new set of informations and their covariance matrices, $e_t^{(i)}$ and $V_t^{(i)}$. The expressions $\delta_i^{-1}\left(e_t^{(i)} - e_t\right)$ and $\delta_i^{-1}\left(V_t^{(i)} - V_t\right)$ are then numerical approximations

to the required derivatives.

The derivatives of V_t and e_t can be evaluated analytically by programming n sets of recursions to run in parallel with the Kalman filter. The i^{th} set of recursions gives the quantities needed to calculate the derivatives of V_t and e_t with respect of the i^{th} element of Ω.

Since

$$e_t = y_t - H_t \mathbf{s}(t|t-1) \qquad t = 1, \ldots, T,$$

the vector of derivatives with respect to Ω_i is

$$\frac{\partial e_t}{\partial \Omega_i} = -H_t \frac{\partial \mathbf{s}(t|t-1)}{\partial \Omega_i} - \frac{\partial H_t}{\partial \Omega_i}\mathbf{s}(t|t-1)$$

and so a recursion is needed to provide $\partial \mathbf{s}(t|t-1)/\partial \Omega_i$. Similarly,

$$\frac{\partial V_t}{\partial \Omega_i} = \frac{\partial H_t}{\partial \Omega_i} P(t|t-1) H_t^T + H_t \frac{\partial P(t|t-1)}{\partial \Omega_i} H_t^T + H_t^T P(t|t-1) \frac{\partial H_t^T}{\partial \Omega_i} + \frac{\partial \Sigma_t \Sigma_t^T}{\partial \Omega_i}$$

and so a recursion is also needed for the derivatives of $P(t|t-1)$.

The recursion for the derivatives of $\mathbf{s}(t|t-1)$ and $P(t|t-1)$ are obtained by differentiating the Kalman filter prediction and updating the state and its covariance matrix. Differentiating the prediction equations, gives

$$\frac{\partial \mathbf{s}(t|t-1)}{\partial \Omega_i} = \frac{\partial \Phi_t}{\partial \Omega_i}\mathbf{s}(\mathbf{t-1}) + \Phi_t \frac{\partial \mathbf{s}(t-1)}{\partial \Omega_i} \tag{4.9}$$

and

$$\frac{\partial P(t|t-1)}{\partial \Omega_i} = \frac{\partial \Phi_t}{\partial \Omega_i} P(t-1)\Phi_t^T + \Phi_t \frac{\partial P(t-1)}{\partial \Omega_i}\Phi_t^T + \Phi_t P(t-1)\frac{\partial \Phi_t^T}{\partial \Omega_i} \tag{4.10}$$

While for the updating equations

$$\begin{aligned}
\frac{\partial \mathbf{s}(t|t)}{\partial \Omega_i} &= \frac{\partial \mathbf{s}(t|t-1)}{\partial \Omega_i} + \frac{\partial P(t|t-1)}{\partial \Omega_i} H_t V_t^{-1} \mathbf{e}_t \\
&\quad + P(t|t-1)\frac{\partial H_t}{\partial \Omega_i} V_t^{-1}\mathbf{e}_t - P(t|t-1)H_t^T V_t^{-1}\frac{\partial V_t}{\partial \Omega_i} V_t^{-1}\mathbf{e}_t \\
&\quad + P(t|t-1)H_t^T V_t^{-1}\frac{\partial \mathbf{e}_t}{\partial \Omega_i}
\end{aligned} \tag{4.11}$$

and

$$\begin{aligned}
\frac{\partial P(t|t)}{\partial \Omega_i} &= \frac{\partial P(t|t-1)}{\partial \Omega_i} - \frac{\partial P(t|t-1)}{\partial \Omega_i} H_t^T V_t^{-1} H_t P(t|t-1) \\
&\quad - P(t|t-1)\frac{\partial H_t}{\partial \Omega_i} V_t^{-1} H_t P(t|t-1) + P(t|t-1)H_t^T \frac{\partial V_t^{-1}}{\partial \Omega_i} H_t P(t|t-1) \\
&\quad - P(t|t-1)H_t^T V_t^{-1}\frac{\partial H_t}{\partial \Omega_i} P(t|t-1) \\
&\quad - P(t|t-1)H_t^T V_t^{-1} H_t \frac{\partial P(t|t-1)}{\partial \Omega_i}
\end{aligned} \tag{4.12}$$

for $t = 1, \ldots, T$. Equations (4.9), (4.10), (4.11) and (4.12) give the required derivatives. The initial values of the derivatives depend on the initial values of the proper Kalman filter.

Chapter 5

Applications

5.1 Growth of Sitka spruce data

Dr Peter Lucas of the Biological Sciences Division at Lancaster University provided Diggle et al (1994) this data on the growth of Sitka spruce trees. Since ozone pollution is common in urban areas, the impact of increased ozone concentrations on tree growth is of considerable interest. Dr Lucas study aim was to evaluate the effect of ozone pollution on the tree growth.

The response variable is log tree size, where size is conventionally measured by the product of the tree height and diameter squared. The original data comprised 79 trees over two growing seasons, however in this example, only the data on the first growing season was used. A total of 54 trees were grown with ozone exposure at 70 ppb in two chambers containing 27 trees each. The remaining trees were grown under controlled conditions in two other chambers containing 12 and 13 trees each.

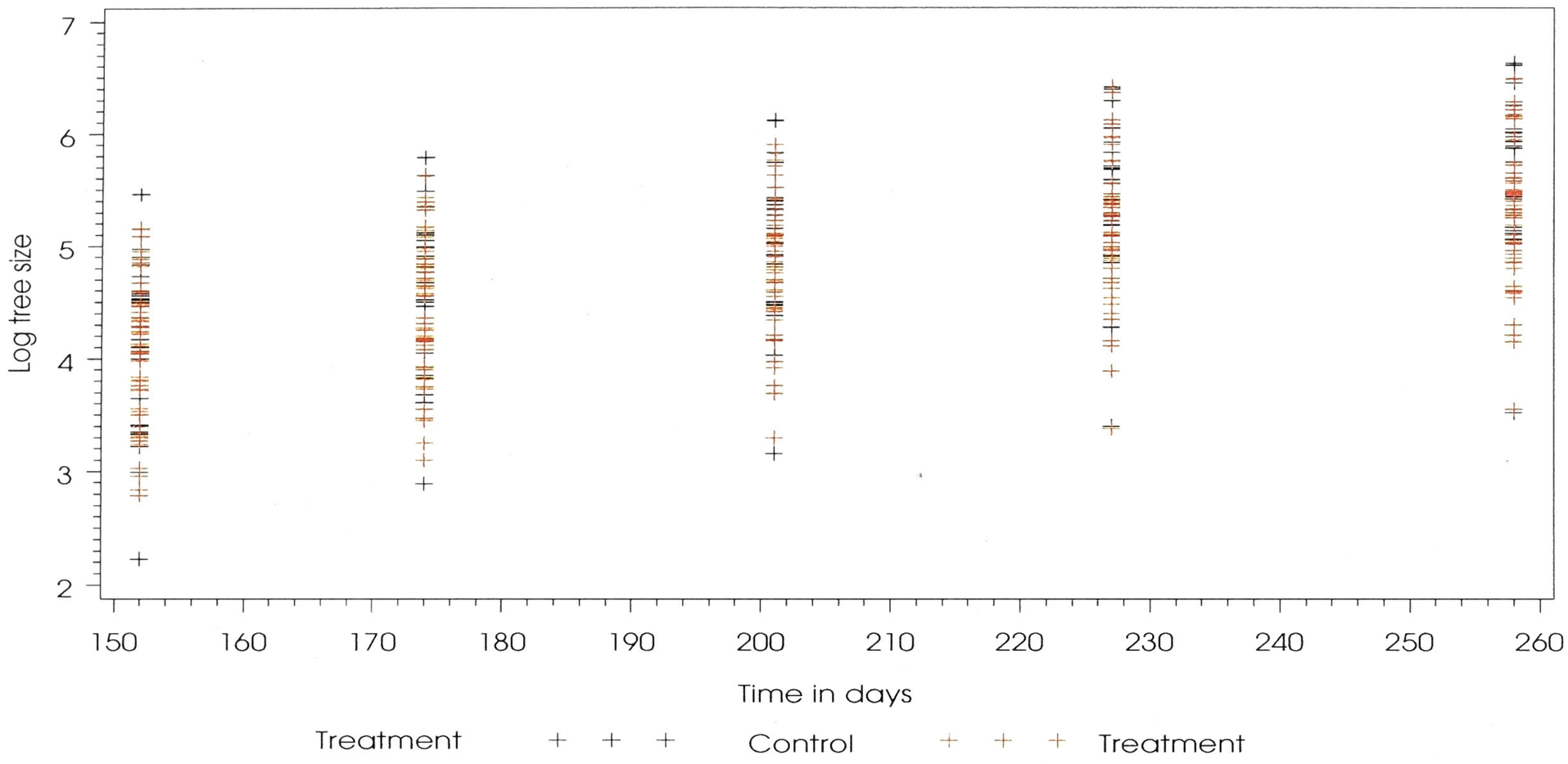

Figure 5.1: Spruce data

For the purpose of simplifying the example, only the first growing season was considered. Also, the observations were considered to be equally spaced in time, even though the time intervals between observations were slightly different.

An appropriate model for this data is:

$$y_{ijt} = E(y_{ijt}) + \epsilon_{ijt}, \tag{5.1}$$

where y_{ijt} is the response of tree i receiving treatment j at time t. The expected value of y_{ijt} can be written as:

$$E(y_{ijt}) = \beta_1 I(t=1) + \beta_2 I(t=2) + \beta_3 I(t=3) + \beta_4 I(t=4) + \beta_5 I(t=5) + \beta_6 I(j=1) \tag{5.2}$$

and the errors ϵ_{ijt} are autocorrelated in the following form:

$$\epsilon_{ijt} = \phi\epsilon_{ijt-1} + \psi_t u_t. \tag{5.3}$$

As was shown in chapter 3, this model can be written in state - space form with a state matrix

$$S = (s_{x_1}, s_{x_2}, s_{x_3}, s_{x_4}, s_{x_5}, s_{x_6}, s_y) \tag{5.4}$$

$$= (x_1, x_2, x_3, x_4, x_5, x_6, y) \tag{5.5}$$

where $(x_1, x_2, x_3, x_4, x_5, x_6, y)$ is a 375×7 matrix of observations, $s_{x_1}, \ldots, s_{x_6}$ are the unobserved state vectors for $x_1, \ldots, x_6$ respectively and s_y is the unobserved vector for y. The rows of $(x_1, x_2, x_3, x_4, x_5, x_6, y)$ are $(x_{1t}, \ldots, x_{6t}, y_t)$ and the rows of the state matrix S are $s_t = (s_{x_{1t}}, \ldots, s_{x_{6t}}, s_{y_t})$. The transition equation is

$$s_t = \phi s_{t-1} + \psi_t u_t. \tag{5.6}$$

The Kalman filter can be initialized with:

$$s(0|0) = (0, 0, 0, 0, 0, 0, 0),$$

$$\mathbf{P}(0|0) = 1,$$

$$\Psi^2 = 1 - \phi^2,$$

$$\mathrm{M} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

and

$$\Delta = 0.$$

After running the Kalman filter as in section (3.2) and optimizing the likelihood function with respect to ϕ we found that $\hat{\phi} = 0.9719$, $\Delta = -914.1843$, $\hat{\sigma^2} = 0.4104$ and

$$\mathrm{M} = \begin{pmatrix} 1425.72 & -1385.66 & -1.54E-13 & -1.71E-29 & -1.9E-45 & 27.38 & -425.18 \\ -1385.66 & 2772.45 & -1385.66 & -1.54E-13 & -1.71E-29 & 0.77 & 53.67 \\ -1.54E-13 & -1385.66 & 2772.45 & -1385.66 & -1.54E-13 & 0.77 & 55.87 \\ -1.71E-29 & -1.54E-13 & -1385.66 & 2772.45 & -1385.66 & 0.77 & 276.22 \\ -1.9E-45 & -1.71E-29 & -1.54E-13 & -1385.66 & 1425.73 & 27.38 & 437.14 \\ 27.39 & 0.77 & 0.77 & 0.77 & 27.38 & 57.08 & 267.83 \\ -425.18 & 53.67 & 55.87 & 276.22 & 437.14 & 267.83 & 2762.71 \end{pmatrix}$$

so

$$(x^{*T}x^*) = \begin{pmatrix} 1425.72 & -1385.66 & -1.54E-13 & -1.71E-29 & -1.9E-45 & 27.38 \\ -1385.66 & 2772.45 & -1385.66 & -1.54E-13 & -1.71E-29 & 0.77 \\ -1.54E-13 & -1385.66 & 2772.45 & -1385.66 & -1.54E-13 & 0.77 \\ -1.71E-29 & -1.54E-13 & -1385.66 & 2772.45 & -1385.66 & 0.77 \\ -1.9E-45 & -1.71E-29 & -1.54E-13 & -1385.66 & 1425.73 & 27.38 \\ 27.39 & 0.77 & 0.77 & 0.77 & 27.38 & 57.08 \end{pmatrix},$$

$$(x^{*T}y^{*}) = \begin{pmatrix} -425.18 \\ 53.67 \\ 55.87 \\ 276.22 \\ 437.14 \\ 267.83 \end{pmatrix},$$

and

$$(y^{*T}y^{*}) = \begin{pmatrix} 2762.71 \end{pmatrix}.$$

The vector of estimated parameters can be calculated as $\hat{\beta} = (x^{*T}x^{*})^{-1}x^{*T}y^{*}$, to yield

$$\hat{\beta} = \begin{pmatrix} 4.2457 \\ 4.6709 \\ 5.0610 \\ 5.4148 \\ 5.5735 \\ -0.2230 \end{pmatrix}.$$

The estimated standard errors, $\hat{s}$ for the parameters estimates $\hat{\beta}$ can be calculated as the squared root of the diagonal of $\hat{\sigma}^2(x^{*T}x^{*})^{-1}$ so

$$\hat{s} = \begin{pmatrix} 0.0158 \\ 0.0158 \\ 0.0158 \\ 0.0158 \\ 0.0158 \\ 0.0227 \end{pmatrix},$$

and the t-statistics for the parameters estimates$\hat{\beta}$ as $\hat{\beta}_k/\hat{s}_k$ for $k = 1, \ldots, 6$. It is possible to construct the table of the estimated parameters and their significance.

$Parameter$	$Estimate$	$StdError$	$t-Stat$
$\hat{\beta}_1$	4.2457	0.0158	33.77
$\hat{\beta}_2$	4.6709	0.0158	37.15
$\hat{\beta}_3$	5.0610	0.0158	40.25
$\hat{\beta}_4$	5.4148	0.0158	43.06
$\hat{\beta}_5$	5.5735	0.0158	44.33
$\hat{\beta}_6$	−0.2230	0.0227	1.48

This shows that the ozone treatment is not statistically significant in affecting the tree growth but time is.

If the same model is fitted ignoring the temporal correlation of the errors, the table of the estimated parameters and their significance levels is

$Parameter$	$Estimate$	$StdError$	$t-Stat$
$\hat{\beta}_1$	4.2376	0.0071	50.43
$\hat{\beta}_2$	4.6628	0.0071	55.49
$\hat{\beta}_3$	5.0529	0.0071	60.13
$\hat{\beta}_4$	5.4067	0.0071	64.34
$\hat{\beta}_5$	5.5655	0.0071	66.23
$\hat{\beta}_6$	−0.2112	0.0046	3.13

The mean parameters in this model are not very different to the mean parameters in the model that takes into account the error correlations. However, the variance is much smaller, making the treatment effect significant.

5.2 Diabetes data

This data contains clinical information about 73 children that developed diabetes at some stage of their lives. It is of interest to see if growth is different for boys and girls, if the diabetes status of their mother affects growth and if the way in which they were diagnosed with diabetes affected their growth.

The children's weights were measured at birth, and then at different time intervals during their childhood. These intervals range from one day when they were young babies to about one year and nine months when they were older. These intervals vary from patient to patient. The number of observations per child range from 12 to 285. The total number of observations was 8646.

The diabetes status of their mother has three values, diabetic, not diabetic and unknown. The way in which they were diagnosed with diabetes also has three values, random, gestational and diagnosed. There were only three children with gestational diabetes and they were all girls, so for the purpose of this analysis only children with diagnosed or random diabetes were considered. Also there were no children with unknown mothers's diabetic status and their own diabetic status diagnosed, so this interaction was not included in this analysis.

Figure (5.2) shows the children weight against their age in months. This graph shows pronounced curvature as is expected in a growth curve, so a log transformation was used. The transformed data is shown in Figure (5.3), the curvature has disappeared and the variation looks fairly constant.

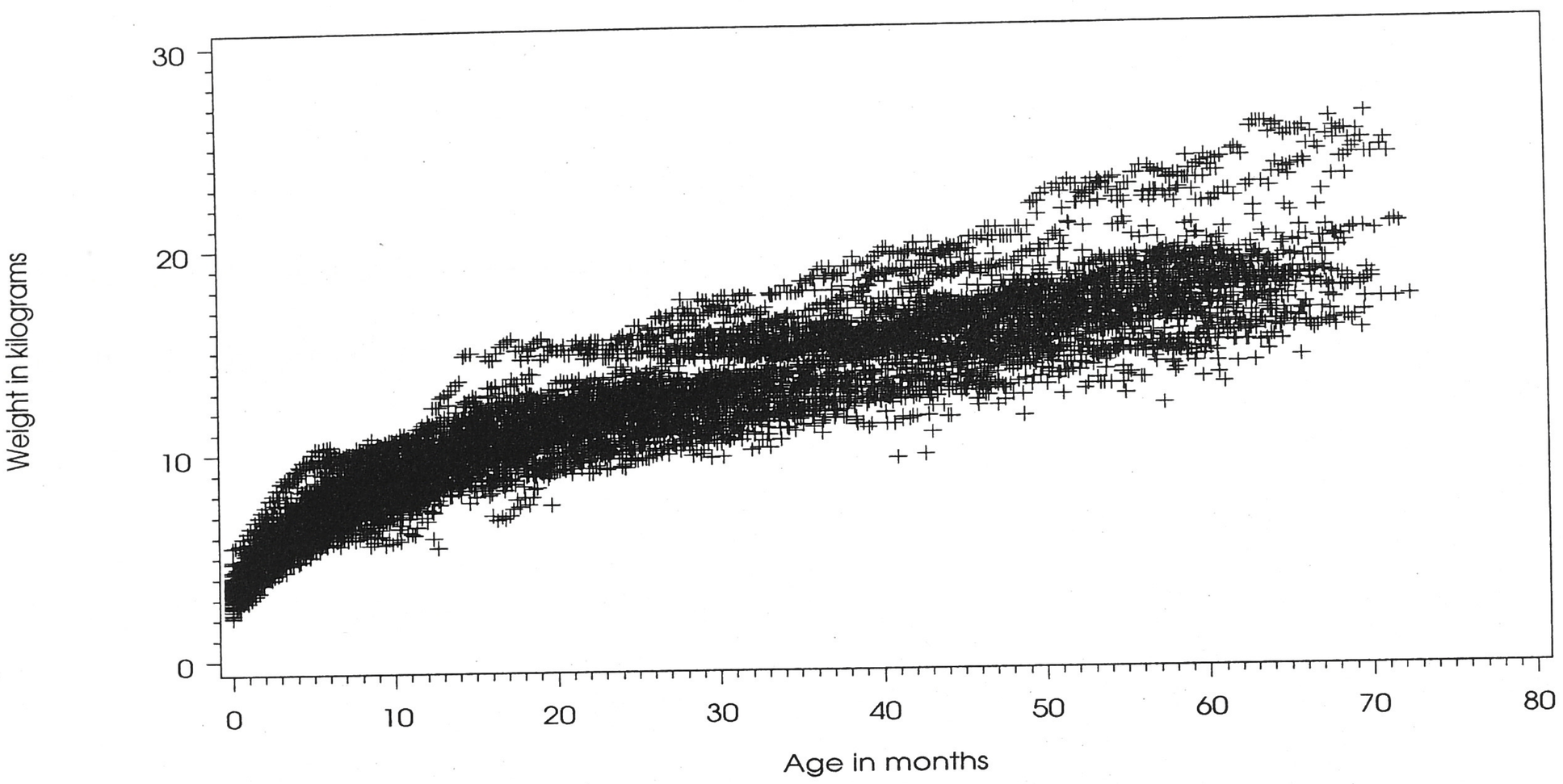

Figure 5.2: Diabetes data

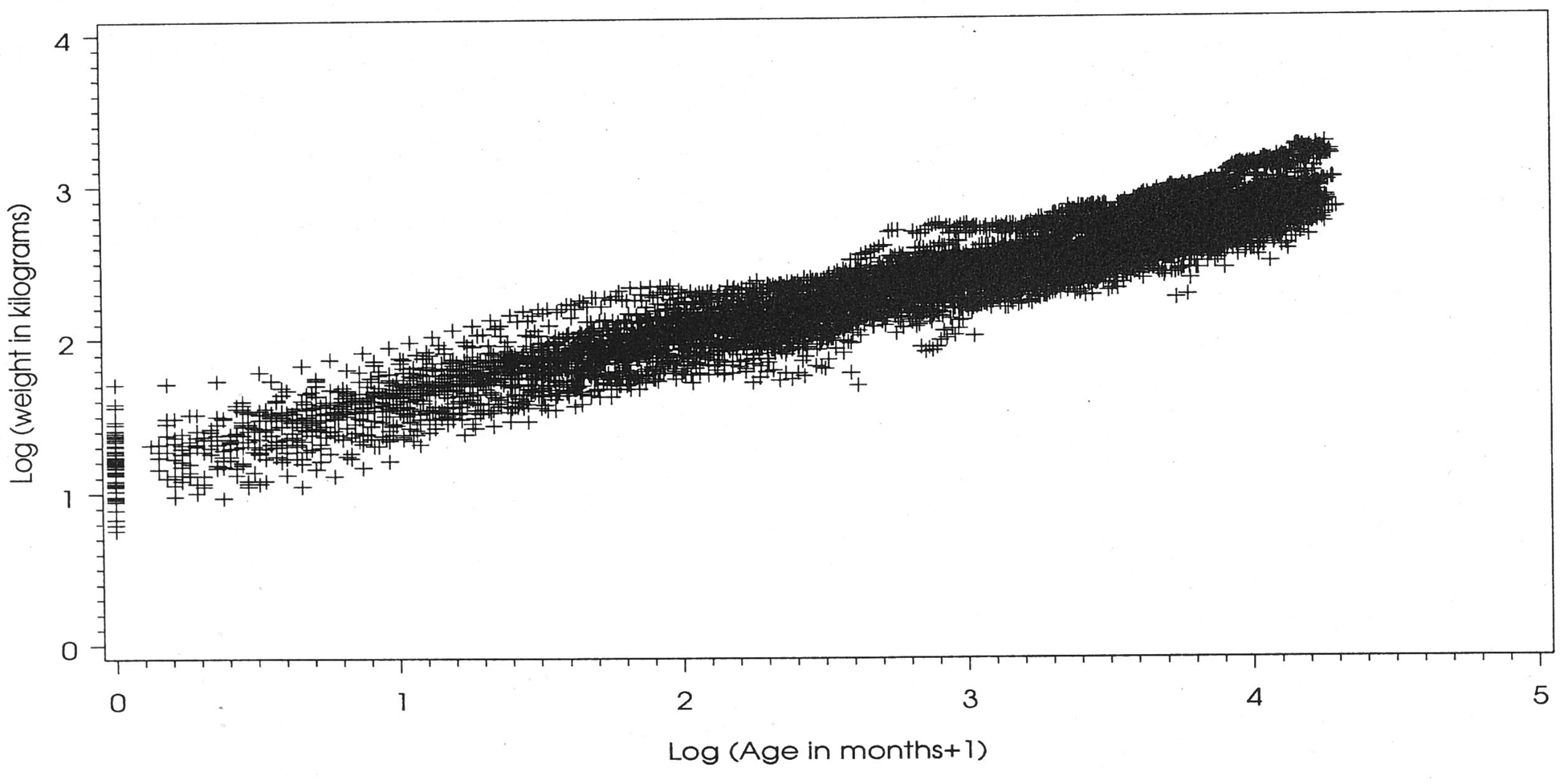

Figure 5.3: Transformed diabetes data

An appropriate model for these data is:

$$y_{ijkmt} = E(y_{ijkmt}) + \epsilon_{ijkmt}, \tag{5.7}$$

where y_{ijkmt} is the $log(weight)$ at time t of child i, with gender j, mother's diabetes status k and diabetes status m and

$$\begin{aligned} E(y_{ijkmt}) &= \mu + \beta_1 x_{ijkmt} + \beta_2 I(j=1) + \beta_3 I(k=1) + \beta_4 I(k=2) \\ &\quad + \beta_5 I(m=1) + \beta_6 x_{ijkmt} * I(j=1) + \beta_7 x_{ijkmt} * I(k=1) \\ &\quad + \beta_8 x_{ijkmt} * I(k=2) + \beta_9 x_{ijkmt} * I(m=1) \\ &\quad + \beta_{10} I(j=1) * I(k=1) + \beta_{11} I(j=1) * I(k=2) + \beta_{12} I(j=1) * I(m=1) \\ &\quad + \beta_{13} I(k=2) * I(m=1), \end{aligned} \tag{5.8}$$

where μ_{ijkmt} is the mean weight at birth of a girl with mother diabetic and diabetes status random, and x_{ijkmt} is the $log(z_{it}+1)$, where z_{it} is the age in months of child i, with gender j, mother's diabetes status k and diabetes status m, at time t. The errors ϵ_{ijkmt} are correlated in the following form:

$$\epsilon_{ijkmt} = e^{-\alpha(z_{it}-z_{it-1})}\epsilon_{ijkmt-1} + \psi_t u_t. \tag{5.9}$$

This model can be written in state-space form with a state matrix

$$S = (s_{x_1}, s_{x_2}, s_{x_3}, s_{x_4}, s_{x_5}, s_{x_6}, s_{x_7}, s_{x_8}, s_{x_9}, s_{x_{10}}, s_{x_{11}}, s_{x_{12}}, s_{x_{13}}, s_{x_{14}}, s_y) \tag{5.10}$$

$$= (x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}, x_{13}, x_{14}, y) \tag{5.11}$$

where $(x_1, \ldots, x_{14}, y)$ is an 8646×15 matrix of observations, $s_{x_1}, \ldots, s_{x_{14}}$ are the unobserved state vectors for $x_1, \ldots, x_{14}$ respectively and s_y is the unobserved vector for y. The rows of $(x_1, \ldots, x_{14}, y)$ are $(x_{1t}, \ldots, x_{14t}, y_t)$ and the rows of the state matrix S are $s_t = (s_{x_{1t}}, \ldots, s_{x_{14t}}, s_{y_t})$. The transition equation is

$$s_t = e^{-\alpha(z_t - z_{t-1})} s_{t-1} + \psi_t u_t. \tag{5.12}$$

To concentrate σ^2 out of the likelihood, the Kalman filter can be initialized with:

$$s(t_1|0) = (0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0),$$

$$\mathbf{P}(t_1|0) = \frac{1}{2\alpha},$$

$$\Psi^2(z_t - z_{t-1}) = \frac{1}{2\alpha}(1 - e^{-2\alpha(z_t - z_{t-1})}),$$

M is a 15×15 zero matrix and $\Delta = 0$.

After running the Kalman filter as described in Section (3.2) and optimizing the likelihood function with respect to α it is found that $\hat{\alpha} = 0.0937$, $\Delta = -9153.6$, $\hat{\sigma}^2 = 0.0035$ and

$$
\mathbf{M} = \begin{pmatrix}
53 & 155 & 25 & 26 & 18 & 10 & 73 & 77 & 51 & 30 & 13 & 10 & 7 & 2 & 126 \\
155 & 624 & 73 & 77 & 51 & 30 & 296 & 312 & 205 & 122 & 38 & 30 & 21 & 7 & 435 \\
25 & 73 & 25 & 13 & 10 & 7 & 73 & 38 & 30 & 21 & 13 & 10 & 7 & 1 & 60 \\
26 & 77 & 13 & 26 & 0 & 8 & 38 & 77 & 0 & 23 & 13 & 0 & 6 & 0 & 63 \\
18 & 51 & 10 & 0 & 18 & 2 & 30 & 0 & 51 & 7 & 0 & 10 & 1 & 2 & 41 \\
10 & 30 & 7 & 8 & 2 & 10 & 21 & 23 & 7 & 30 & 6 & 1 & 7 & 2 & 24 \\
73 & 296 & 73 & 38 & 30 & 21 & 296 & 155 & 121 & 85 & 38 & 30 & 21 & 2 & 209 \\
77 & 312 & 38 & 77 & 0 & 23 & 155 & 312 & 0 & 94 & 38 & 0 & 19 & 0 & 219 \\
51 & 205 & 30 & 0 & 51 & 7 & 121 & 0 & 205 & 28 & 0 & 30 & 2 & 7 & 142 \\
30 & 122 & 21 & 23 & 7 & 30 & 85 & 94 & 28 & 122 & 19 & 2 & 21 & 7 & 85 \\
13 & 38 & 13 & 13 & 0 & 6 & 38 & 38 & 0 & 19 & 13 & 0 & 6 & 0 & 32 \\
10 & 30 & 10 & 0 & 10 & 1 & 30 & 0 & 30 & 2 & 0 & 10 & 1 & 1 & 24 \\
7 & 21 & 7 & 6 & 1 & 7 & 21 & 19 & 2 & 21 & 6 & 1 & 7 & 1 & 17 \\
2 & 7 & 1 & 0 & 2 & 2 & 2 & 0 & 7 & 7 & 0 & 1 & 1 & 2 & 5 \\
126 & 435 & 60 & 63 & 41 & 24 & 209 & 219 & 142 & 85 & 32 & 24 & 17 & 5 & 355
\end{pmatrix},
$$

SO

$$(x^{*T}x^*) = \begin{pmatrix}
53 & 155 & 25 & 26 & 18 & 10 & 73 & 77 & 51 & 30 & 13 & 10 & 7 & 2 \\
155 & 624 & 73 & 77 & 51 & 30 & 296 & 312 & 205 & 122 & 38 & 30 & 21 & 7 \\
25 & 73 & 25 & 13 & 10 & 7 & 73 & 38 & 30 & 21 & 13 & 10 & 7 & 1 \\
26 & 77 & 13 & 26 & 0 & 8 & 38 & 77 & 0 & 23 & 13 & 0 & 6 & 0 \\
18 & 51 & 10 & 0 & 18 & 2 & 30 & 0 & 51 & 7 & 0 & 10 & 1 & 2 \\
10 & 30 & 7 & 8 & 2 & 10 & 21 & 23 & 7 & 30 & 6 & 1 & 7 & 2 \\
73 & 296 & 73 & 38 & 30 & 21 & 296 & 155 & 121 & 85 & 38 & 30 & 21 & 2 \\
77 & 312 & 38 & 77 & 0 & 23 & 155 & 312 & 0 & 94 & 38 & 0 & 19 & 0 \\
51 & 205 & 30 & 0 & 51 & 7 & 121 & 0 & 205 & 28 & 0 & 30 & 2 & 7 \\
30 & 122 & 21 & 23 & 7 & 30 & 85 & 94 & 28 & 122 & 19 & 2 & 21 & 7 \\
13 & 38 & 13 & 13 & 0 & 6 & 38 & 38 & 0 & 19 & 13 & 0 & 6 & 0 \\
10 & 30 & 10 & 0 & 10 & 1 & 30 & 0 & 30 & 2 & 0 & 10 & 1 & 1 \\
7 & 21 & 7 & 6 & 1 & 7 & 21 & 19 & 2 & 21 & 6 & 1 & 7 & 1 \\
2 & 7 & 1 & 0 & 2 & 2 & 2 & 0 & 7 & 7 & 0 & 1 & 1 & 2
\end{pmatrix},$$

$$(x^{*T}y^*) = \begin{pmatrix} 126 \\ 435 \\ 60 \\ 63 \\ 41 \\ 24 \\ 209 \\ 219 \\ 142 \\ 85 \\ 32 \\ 24 \\ 17 \\ 5 \end{pmatrix},$$

and

$$(y^{*T}y^*) = \begin{pmatrix} 355 \end{pmatrix}.$$

So the parameters vector can be calculated as $\hat{\beta} = (x^{*T}x^*)^{-1}x^{*T}y^*$, then

$$\hat{\beta} = \begin{pmatrix} 1.2002 \\ 0.3963 \\ 0.0190 \\ -0.0160 \\ -0.0124 \\ 0.0227 \\ 0.0129 \\ 0.0002 \\ -0.0043 \\ -0.0089 \\ 0.0051 \\ -0.0257 \\ 0.0397 \\ -0.0262 \end{pmatrix}.$$

The estimated standard errors, $\hat{s}$ for the parameters estimates $\hat{\beta}$ can be calculated as the squared root of the diagonal of $\hat{\sigma}^2(x^{*T}x^*)^{-1}$ so

$$\hat{s} = \begin{pmatrix} 0.0158 \\ 0.0014 \\ 0.0001 \\ 0.0035 \\ 0.0022 \\ 0.0027 \\ 0.0033 \\ 0.0001 \\ 0.0002 \\ 0.0002 \\ 0.0002 \\ 0.0035 \\ 0.0036 \\ 0.0025 \\ 0.0029 \end{pmatrix},$$

and the t-statistics for the parameters estimates $\hat{\beta}$ as $\hat{\beta}_k/\hat{s}_k$ for $k = 1, \ldots, 14$. So it is possible to construct the table of the estimated parameters and their significance.

$Parameter$	$Estimate$	$StdError$	$t-Stat$
$\hat{\mu}$	1.2002	0.0380	31.62
$log(ageinmonths+1)$	0.3963	0.0109	36.34
$Gender, boy$	0.0190	0.0596	0.32
$Mother, diabetic, unknown$	-0.0160	0.0473	0.34
$Mother, not-diabetic$	-0.0124	0.0516	0.24
$Diabetes-diagnosed$	0.0227	0.0582	0.39
$log(ageinmonths+1)*Gender$	0.0129	0.0097	1.33
$log(ageinmonths+1)*Mother, diabetic, unknown$	0.0002	0.0133	0.01
$log(ageinmonths+1)*Mother, not-diabetic$	-0.0043	0.0140	0.31
$log(ageinmonths+1)*Diabetes-diagnosed$	-0.0089	0.0124	0.71
$Gender*Mother, diabetic, unknown$	0.0051	0.0589	0.09
$Gender*Mother, not-diabetic$	-0.0257	0.0601	0.43
$Gender*Diabetes-diagnosed$	0.0397	0.0502	0.79
$Mother, not-diabetic*Diabetes-diagnosed$	-0.0262	0.0537	0.49

Figure (5.4) shows the fitted values against the residuals. This graph does not seem to present major problems, the residuals look randomly spread and the variability is fairly constant along time.

The only parameter that is significant is the one related to age. Note that if state-space models were not used in this example it would have been necessary to invert matrices of dimensions up to 285×285, and using the Kalman filter avoids this.

Jones and Boadi (1991) pointed out that when continuous serial correlation is included in the model, it is necessary to scale the time intervals. Time could be measured in years, months, weeks or days, and the analysis is invariant with respect of the chosen

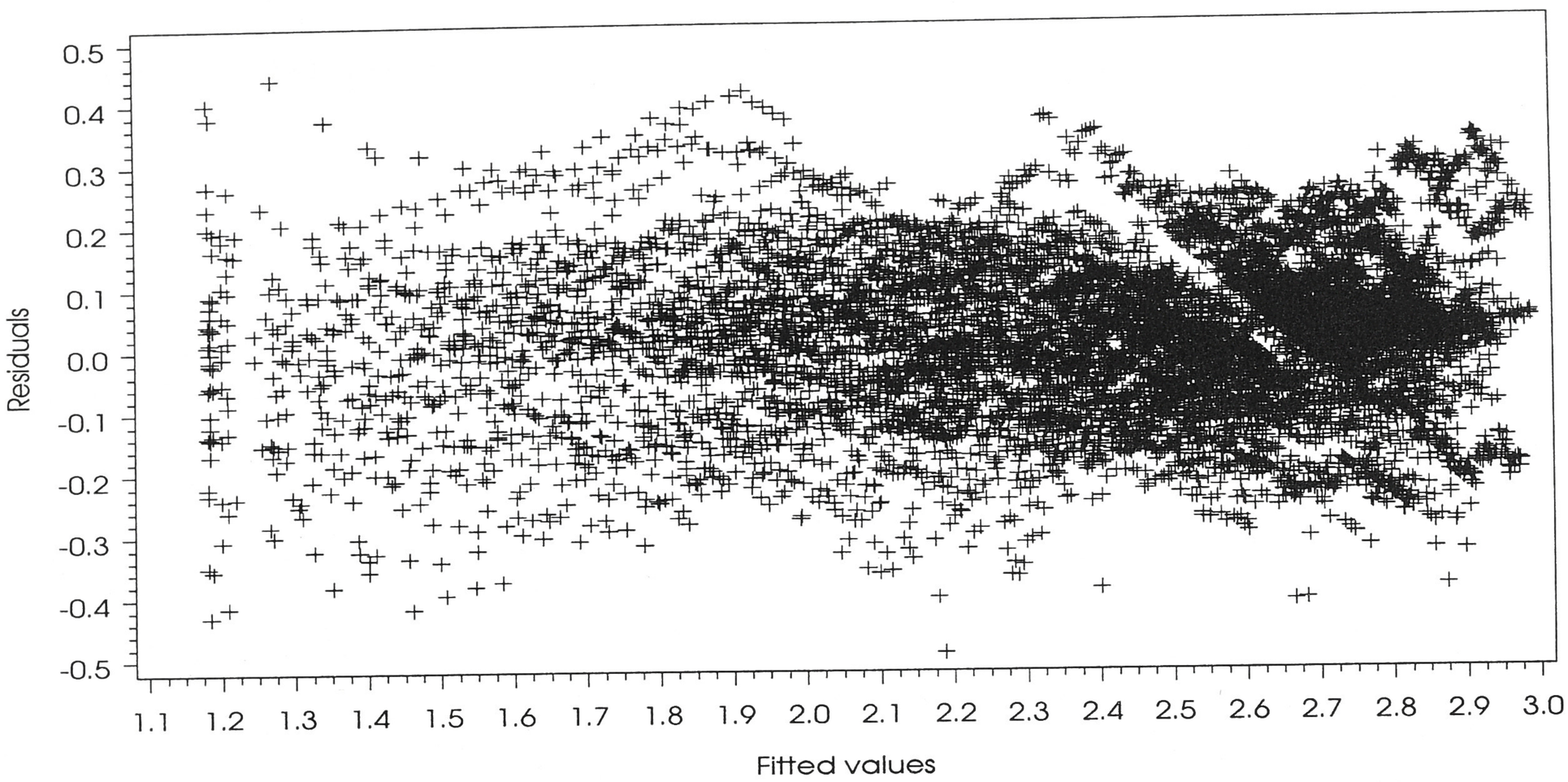

Figure 5.4: Diabetes data: fitted values against residuals

unit of time. However, since exponential functions are involved, some combinations of unit of time and guesses at non linear parameters can cause overflow or underflow problems. For example, if the time unit is days, the time interval for the observations that were one year and nine months apart is 640 days, when considering values of $\alpha = 0.5$ or $\alpha = 1$, the program calculates e^{-320} or e^{-640} respectively, which could underflow. Choosing the time unit in this example was particularly problematic, since the time intervals vary from one day to one year and nine months. Given this large variation the intermediate unit of months worked best.

It is also important to notice that even though there are statistical packages that have functions available to fit these type of models when the observations are equally spaced in time, so the errors have a discrete time structure, there are none for when the errors are unequally spaced in time so have a continuous time structure.

If this same model is fitted ignoring the temporal correlation of the errors, the table of the estimated parameters and their significance levels is

$Parameter$	$Estimate$	$StdError$	$t-Stat$
$\hat{\mu}$	1.16	0.0103	113.51
$log(ageinmonths+1)$	0.4123	0.0003	127.86
$Gender, boy$	0.0080	0.0121	0.66
$Mother, diabetic, unknown$	−0.0288	0.01267	2.26
$Mother, not-diabetic$	0.0519	0.0138	3.75
$Diabetes-diagnosed$	−0.0248	0.0122	2.03
$log(ageinmonths+1)*Gender$	0.0052	0.0029	1.75
$log(ageinmonths+1)*Mother, diabetic, unknown$	0.0058	0.0040	1.44
$log(ageinmonths+1)*Mother, not-diabetic$	−0.0291	0.0043	6.78
$log(ageinmonths+1)*Diabetes-diagnosed$	−0.0128	0.0034	3.80
$Gender*Mother, diabetic, unknown$	0.0591	0.0098	6.03
$Gender*Mother, not-diabetic$	0.0266	0.0105	2.54
$Gender*Diabetes-diagnosed$	0.0473	0.0071	6.63
$Mother, not-diabetic*Diabetes-diagnosed$	0.0244	0.0073	3.33

In this model six of the interactions are significant. As in the Sitka Spruce example, the variance in this model is much smaller than the variance in the model that takes into account the error correlations making many parameters significant.

Chapter 6

Conclusion

Many statistical models for longitudinal data can be written in state-space form. Some simple models like the autoregressive model are almost already written in state-space form, others like a simple linear regression model are not so intuitive. Also some models can be written in state space form in more than one way. Through out this thesis, only one state-space form for each model was studied.

Once the model is written in state-space form, the Kalman filter can be applied to it to obtain the likelihood function. A key issue in applying the Kalman filter is setting the initial values of the state vector $s(0|0)$ and its covariance matrix $P(0|0)$. These initial values are given by the mean and covariance matrix of the unconditional distribution of the state vector.

Section (2.3) showed how an autoregressive model can be written in state-space form and how the Kalman filter is run to obtain this model's likelihood function. It was also shown, that to concentrate out of the likelihood Ψ^2 in an AR(1) model, $P(0|0)$ can be

set equal to either 1 or $1/(1-\phi^1)^2$ for the purpose of running the Kalman filter.

In section (2.4) the continuous AR(1) process, the CAR(1) process, was studied and in section (2.5) this model was written in state-space form and its likelihood function was found using the Kalman filter.

The simple linear regression model was also studied. Specifically, how the Kalman filter can be used to write its likelihood function, and how the Kalman filter effectively performs the Cholesky decomposition. In the case of linear regression models, it is necessary to run the Kalman filter for the dependent variable, as well as for each of the independent variables. Also, the state vector and its covariance have to be re-initialized for each subject. In a simple linear regression model with errors correlated with an AR(1) structure, the state covariance $P(0|0)$ should be set equal to one to obtain the Cholesky decomposition.

Then the Laird-Ware model was studied. This is a very general mixed model for longitudinal data. This model is very flexible because different subjects can have different numbers of observations as well as different observation times. The Laird-Ware model can be put in state-space form in a similar way as a regression model but including the random effects.

Usually the aim of obtaining the likelihood function of a model is to estimate the unknown parameters by Maximum Likelihood estimation. To do this it is necessary to calculate the first and second derivatives of the likelihood function. Section (4.4) showed how this could be done using the Kalman filter.

The growth of Sitka spruce trees data was analyzed as a simple first example. In this example, each tree had the same number of observations and those observations were equally spaced in time. The aim of this study was to evaluate the effect of ozone pollution on the Sitka spruce trees. The original data comprised 79 trees over two growing seasons, but to simplify this example only the first growing season was used. A total of 54 trees were grown with ozone exposure at 70 ppb in two chambers containing 27 trees each. The remaining trees were grown under controlled conditions in two chambers containing 12 and 13 trees each. The tree size was measured five times roughly equally spaced along the growing season.

After fitting a model with errors correlated in time with an AR(1) structure, it was found that this correlation was very important with a correlation coefficient $\hat{\phi} = 0.9719$ and the ozone treatment was not statistically significant at the 5 per cent level. To see how the temporal correlation affected the parameters, the same model was fitted to the data, but treating the errors as independent. In this case, the mean parameters were fairly similar to the mean parameters in the model with correlated errors, however the variance was much smaller making the treatment effect significant.

Another example, the diabetes data, was analyzed in section (5.2). These data contains clinical information about 73 children that developed diabetes at some stage of their lives. It was of interest to see if growth was different for boys and girls, if the diabetes status of their mother affected growth and if the way in which they were diagnosed with diabetes affected their growth.

The children's weights were measured at birth, and then, at different time intervals during their childhood. These intervals ranged from one day when they were young babies to about one year and nine months when they were older. These intervals and the number of observations varied from child to child. The number of observations per child ranged from 12 to 285 and the total number of observations was 8646.

Since observations were made at different time intervals, it was appropriate to fit a model with a continuous time correlation, in this case a CAR(1) model was used in the error structure. Since continuous serial correlation was included in the model, it was necessary to scale the time intervals. Time could have been measured in years, months, weeks or days, and the analysis would have been invariant with respect of the chosen unit of time. However, since exponential functions were involved, some combinations of unit of time and guesses at non linear parameters could cause overflow or underflow problems. Choosing the time unit in this example was particularly problematic, since the time intervals vary from one day to one year and nine months. Given this large variation the intermediate unit of months worked best. The correlation parameter α was estimated to be 0.0937.

In model (5.7), only the parameter related to age was significant. However when the same model but with independent errors was fitted, six of the interactions were significant. As in the Sitka Spruce example, the variance in the model with independent errors was much smaller than the variance in the model that took into account the error correlation making many parameters significant.

Both examples showed that error correlation structure cannot be ignored. In the Sitka spruce example, model (5.1), could be fitted using weighted least squares for which

there are functions readily available in several statistical packages. However, that is not the case for the diabetes example. Even though there are statistical packages that have functions available to fit these type of models when the observations are equally spaced in time, so the errors have a discrete time structure, there are none for when the errors are unequally spaced in time and have a continuous time structure. State-space models and the Kalman filter allow an easy way to fit these much more flexible and complicated models without requiring us to invert large matrices whenever the number of observations for a subject is large. For example, using the Kalman filter in the diabetes example avoided having to invert matrices of dimension up to 285×285.

Bibliography

[1] Box, G. E. P. and Jenkins, G. M. (1976), Time series analysis, forecasting and control. Revised edition, Holden - Day, San Francisco

[2] Chi, E. M. and Reinsel, G. C. (1989), Models for longitudinal data with random effects and AR(1) errors. J. Am. Statist. Assoc., 84, pp. 452-459

[3] Diggle, P. J. (1988), An approach to the analysis of repeated measurements. Biometrics, 44, pp. 959-71

[4] Diggle, P. J. (1990), Time series: A biostatistical introduction. Oxford University Press, Oxford

[5] Diggle, P. J., Liang, K. and Zeger, S. L. (1994), Analysis of longitudinal data. Oxford University Press, Oxford

[6] Duncan, D. B. and Horn, S. D. (1972), Linear dynamic recursive estimation from the viewpoint of regression analysis. J. Am. Statist. Assoc., 67, 815-821

[7] Durbin, J. and Koopman S. J. (2001), Time series analysis by state space methods. Oxford University Press, Oxford

[8] Harvey, A. C. (1989), Forecasting, structural time series models and the Kalman filter. Cambridge University Press, Cambridge

[9] Harville, D. A. (1974), Bayesian inference for variance components using only error contrasts. Biometrika, 61, pp. 383-385

[10] Harville, D. A. (1976), Extensions of the Gauss - Markov theorem to include the estimation of random effects. Ann, Statist., 4, pp. 384-395

[11] Harville, D. A. (1977), Maximum likelihood approaches to variance component estimation and to related problems. J. Am. Statist. Assosc., 72, pp. 320-340

[12] Jones, R. H. (1981), Fitting continuous time autoregressions to discrete data. Applied Time Series Analysis II (D. F. Findley, editor), Academic Press, pp. 651-682

[13] Jones, R. H. (1985), Time series analysis with unequally spaced data. Handbook of Statistics, Vol. 5: Time series in the time domain (E. J. Hannan, P.R. Krishnaiah and M. M. Rao, Eds.), North - Holland, pp. 157-177

[14] Jones R. H.(1993), Longitudinal Data with Serial Correlation: A State-Space Approach. London: Chapman and Hall

[15] Jones, R. H. and Boadi - Boateng, F. (1991), Unequally spaced longitudinal data wit AR(1) serial correlation. Biometrics, 47, pp. 161-175

[16] Kalman, R. E. (1960), A new approach to linear fitting and prediction problems. Trans ASME J. Basic Eng., 82D, pp. 35-45

[17] Laird, N. M. and Ware, J. H. (1982), Random effects models for longitudinal data. Biometrics, 38, pp. 963-974

[18] Schweppe, F. C. (1965), Evaluation of likelihood functions for Gaussian signals. IEEE Trans. Inform. Theory, 11, pp. 61-70